Topic Modelling and Latent Dirichlet Allocation

Stephen Clark
(with thanks to Mark Gales for some of the slides)

Lent 2013

Machine Learning for Language Processing: Lecture 7

MPhil in Advanced Computer Science

Introduction to Probabilistic Topic Models

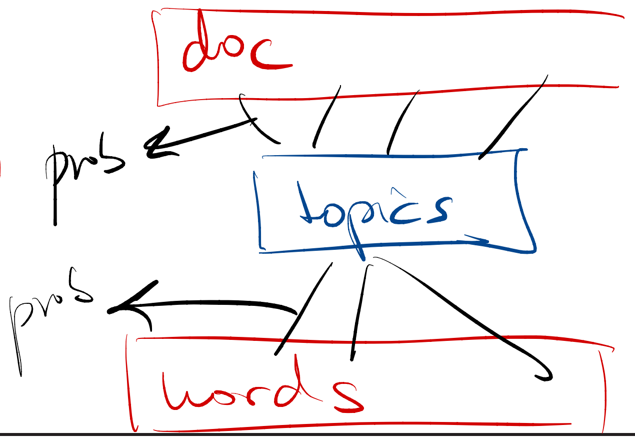
use LDA

latent Dirichlet analysis

- We want to find *themes* (or *topics*) in documents
  - useful for e.g. search or browsing
- We don't want to do supervised topic classification
  - rather not fix topics in advance nor do manual annotation
- Need an approach which automatically teases out the topics
- This is essentially a *clustering* problem - can think of both words and documents as being clustered

Doc = distribution(topics)

Topic = distribution (words)



Key Assumptions behind the LDA Topic Model

- Documents exhibit multiple topics (but typically not many)
- LDA is a probabilistic model with a corresponding *generative process*
  - each document is assumed to be generated by this (simple) process
- A *topic* is a distribution over a fixed vocabulary
  - these topics are assumed to be generated first, before the documents
- Only the number of topics is specified in advance

The Generative Process

To generate a document:

1. Randomly choose a distribution over topics

2. For each word in the document

   a. randomly choose a topic from the distribution over topics
   b. randomly choose a word from the corresponding topic (distribution over the vocabulary)

- Note that we need a distribution over a distribution (for step 1)

- Note that words are generated independently of other words (unigram bag-of-words model)

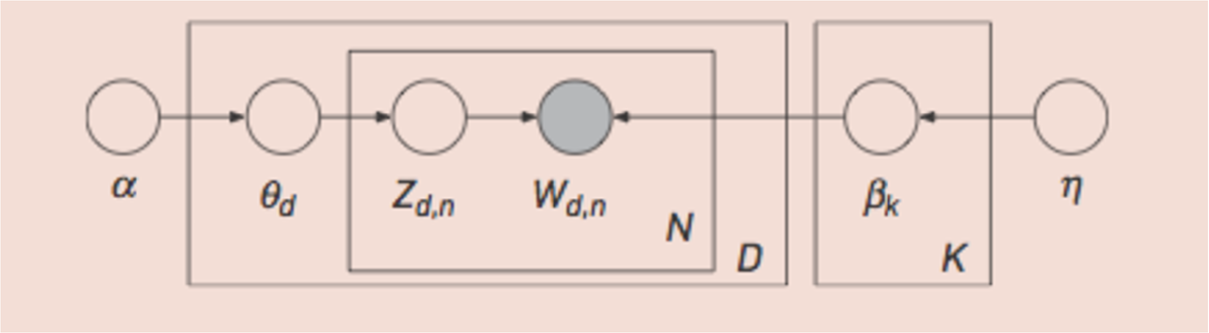
The Generative Process more Formally

- Some notation:
  - $\beta_{1:K}$ are the topics where each β_k is a distribution over the vocabulary
  - θ_d are the topic proportions for document d
  - $\theta_{d,k}$ is the topic proportion for topic k in document d
  - z_d are the topic assignments for document d
  - $z_{d,n}$ is the topic assignment for word n in document d
  - w_d are the observed words for document d

- The joint distribution (of the hidden and observed variables):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K}, z_{d,n})$$

Plate Diagram of the Graphical Model



- Note that only the words are observed (shaded)
- α and η are the parameters of the respective dirichlet distributions (more later)
- Note that the topics are generated (not shown in earlier pseudo code)
- Plates indicate repetition

Picture from Blei 2012

Multinomial Distribution

$$(x+y)^n = \sum \binom{n}{k} x^k y^{n-k}$$

$$\frac{n!}{k!\,(n-k)!}$$

- **Multinomial** distribution: $x_i \in \{0, \ldots, n\}$

vs binomial

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{d} x_i!} \prod_{i=1}^{d} \theta_i^{x_i}, \qquad n = \sum_{i=1}^{d} x_i, \quad \sum_{i=1}^{d} \theta_i = 1, \quad \theta_i \geq 0$$

- When $n = 1$ the multinomial distribution simplifies to

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}, \quad \sum_{i=1}^{d} \theta_i = 1, \quad \theta_i \geq 0$$

  - a unigram language model with 1-of-V coding ($d = V$ the vocabulary size)
  - x_i indicates word i of the vocabulary observed, $x_i = \begin{cases} 1, & \text{word } i \text{ observed} \\ 0, & \text{otherwise} \end{cases}$
  - $\theta_i = P(w_i)$ the probability that word i is seen

The Dirichlet Distribution

- **Dirichlet** (continuous) distribution with parameters $\boldsymbol{\alpha}$

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}; \quad \text{for "observations"}: \sum_{i=1}^{d} x_i = 1, \quad x_i \geq 0$$

- $\Gamma()$ is the Gamma distribution

≃ factorial continous.

- Conjugate prior to the multinomial distribution
  (form of posterior $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ is the same as the prior $p(\boldsymbol{\theta}|\mathcal{M})$)

multinom

p = conjugate (q)

evidence

prior — Dirichle

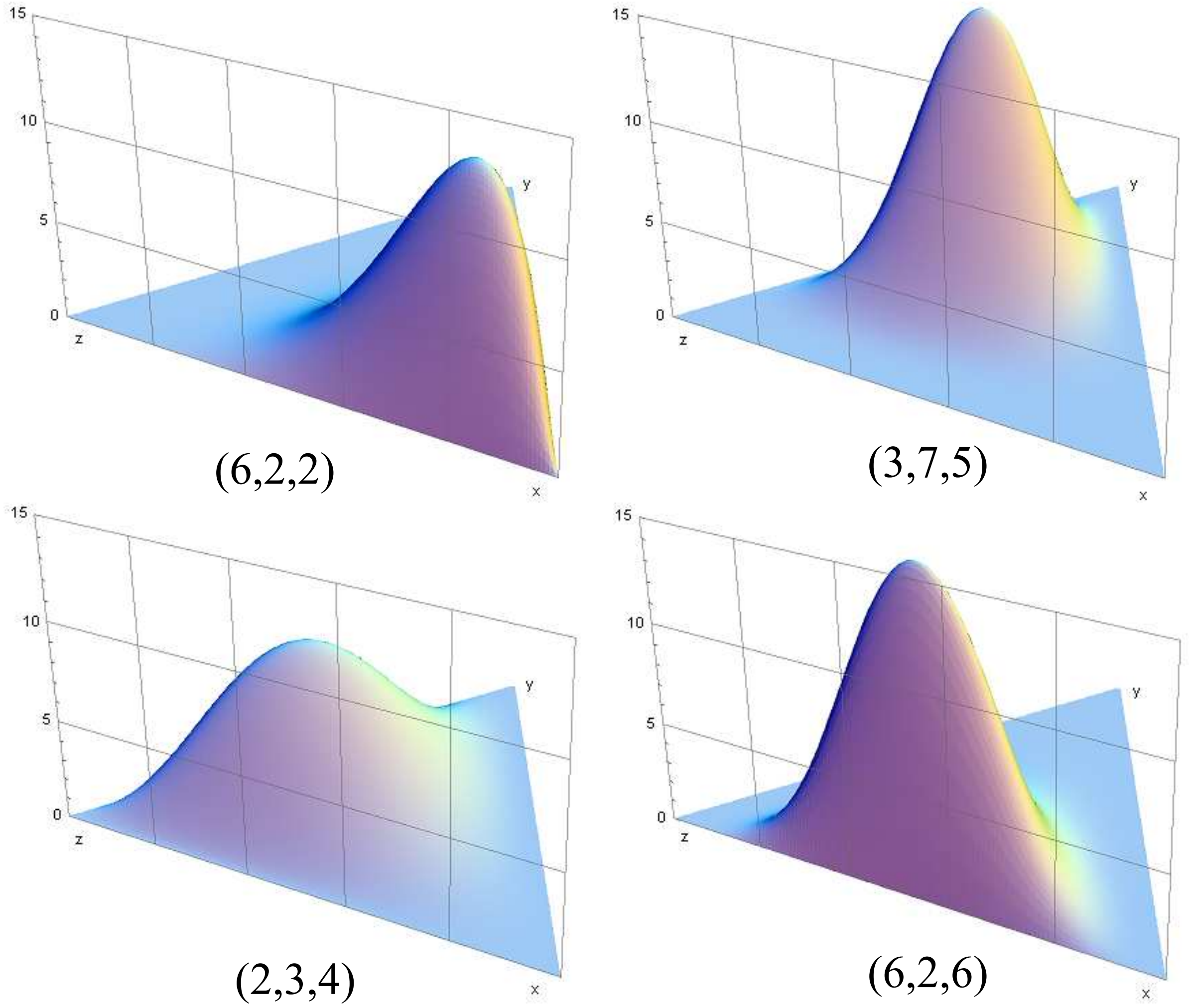
posterior

Dirichlett

$$p(x|Y) = \frac{p(Y|x) \cdot p(x)}{p(Y)}$$

norma

Dirichlet Distribution Example



(6,2,2)

(3,7,5)

(2,3,4)

(6,2,6)

- Parameters: $(\alpha_1, \alpha_2, \alpha_3)$

Parameter Estimation

- Main variables of interest:
  - β_k: distribution over vocabulary for topic k
  - $\theta_{d,k}$: topic proportion for topic k in document d

original approach

- Could try and get these directly, eg using EM (Hoffmann, 1999), but this approach not very successful

- One common technique is to estimate the posterior of the word-topic assignments, given the observed words, directly (whilst marginalizing out β and θ)

Gibbs Sampling

- Gibbs sampling is an example of a Markov Chain Monte Carlo (MCMC) technique
- Markov chain in this instance means that we sample from each variable one at a time, keeping the current values of the other variables fixed

Posterior Estimate

- The Gibbs sampler produces the following estimate, where, following Steyvers and Griffiths:
  - z_i is the topic assigned to the ith token in the whole collection;
  - d_i is the document containing the ith token;
  - w_i is the word type of the ith token;
  - $\mathbf{z}_{-i}$ is the set of topic assignments of all other tokens;
  - $\cdot$ is any remaining information such as the α and η hyperparameters:

$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

where $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are matrices of counts (word-topic and document-topic)

Posterior Estimates of β and θ

$$\beta_{ij} = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta} \quad \theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

- Using the count matrices as before, where β_{ij} is the probability of word type i for topic j, and θ_{dj} is the proportion of topic j in document d

References

- David Blei's webpage is a good place to start

- A good introductory paper: D. Blei. Probabilistic topic models. Communications of the ACM, 55(4):7784, 2012.

- Introduction to Gibbs sampling for LDA: Steyvers, M., Griffiths, T. Probabilistic topic models. Latent Semantic Analysis: A Road to Meaning. T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, eds. Lawrence Erlbaum, 2006.

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

1. Introduction

In this paper we consider the problem of modeling text corpora and other collections of discrete data. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

Significant progress has been made on this problem by researchers in the field of information retrieval (IR) (Baeza-Yates and Ribeiro-Neto, 1999). The basic methodology proposed by IR researchers for text corpora—a methodology successfully deployed in modern Internet search engines—reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts. In the popular *tf-idf* scheme (Salton and McGill, 1983), a basic vocabulary of "words" or "terms" is chosen, and, for each document in the corpus, a count is formed of the number of occurrences of each word. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a word in the entire corpus (generally on a log scale, and again suitably normalized). The end result is a term-by-document matrix X whose columns contain the *tf-idf* values for each of the documents in the corpus. Thus the *tf-idf* scheme reduces documents of arbitrary length to fixed-length lists of numbers.

While the *tf-idf* reduction has some appealing features—notably in its basic identification of sets of words that are discriminative for documents in the collection—the approach also provides a relatively small amount of reduction in description length and reveals little in the way of inter- or intra-document statistical structure. To address these shortcomings, IR researchers have proposed several other dimensionality reduction techniques, most notably *latent semantic indexing (LSI)* (Deerwester et al., 1990). LSI uses a singular value decomposition of the X matrix to identify a linear subspace in the space of *tf-idf* features that captures most of the variance in the collection. This approach can achieve significant compression in large collections. Furthermore, Deerwester et al. argue that the derived features of LSI, which are linear combinations of the original *tf-idf* features, can capture some aspects of basic linguistic notions such as synonymy and polysemy.

To substantiate the claims regarding LSI, and to study its relative strengths and weaknesses, it is useful to develop a generative probabilistic model of text corpora and to study the ability of LSI to recover aspects of the generative model from data (Papadimitriou et al., 1998). Given a generative model of text, however, it is not clear why one should adopt the LSI methodology—one can attempt to proceed more directly, fitting the model to data using maximum likelihood or Bayesian methods.

A significant step forward in this regard was made by Hofmann (1999), who presented the *probabilistic LSI (pLSI)* model, also known as the *aspect model*, as an alternative to LSI. The pLSI approach, which we describe in detail in Section 4.3, models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics." Thus each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics. This distribution is the "reduced description" associated with the document.

While Hofmann's work is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents. In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set.

To see how to proceed beyond pLSI, let us consider the fundamental probabilistic assumptions underlying the class of dimensionality reduction methods that includes LSI and pLSI. All of these methods are based on the "bag-of-words" assumption—that the order of words in a document can be neglected. In the language of probability theory, this is an assumption of *exchangeability* for the words in a document (Aldous, 1985). Moreover, although less often stated formally, these methods also assume that documents are exchangeable; the specific ordering of the documents in a corpus can also be neglected.

A classic representation theorem due to de Finetti (1990) establishes that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture. Thus, if we wish to consider exchangeable representations for documents and words, we need to consider mixture models that capture the exchangeability of both words and documents.

This line of thinking leads to the *latent Dirichlet allocation (LDA)* model that we present in the current paper.

It is important to emphasize that an assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed. Rather, exchangeability essentially can be interpreted as meaning "*conditionally* independent and identically distributed," where the conditioning is with respect to an underlying latent parameter of a probability distribution. Conditionally, the joint distribution of the random variables is simple and factored while marginally over the latent parameter, the joint distribution can be quite complex. Thus, while an assumption of exchangeability is clearly a major simplifying assumption in the domain of text modeling, and its principal justification is that it leads to methods that are computationally efficient, the exchangeability assumptions do not necessarily lead to methods that are restricted to simple frequency counts or linear operations. We aim to demonstrate in the current paper that, by taking the de Finetti theorem seriously, we can capture significant intra-document statistical structure via the mixing distribution.

It is also worth noting that there are a large number of generalizations of the basic notion of exchangeability, including various forms of partial exchangeability, and that representation theorems are available for these cases as well (Diaconis, 1988). Thus, while the work that we discuss in the current paper focuses on simple "bag-of-words" models, which lead to mixture distributions for single words (unigrams), our methods are also applicable to richer models that involve mixtures for larger structural units such as n-grams or paragraphs.

The paper is organized as follows. In Section 2 we introduce basic notation and terminology. The LDA model is presented in Section 3 and is compared to related latent variable models in Section 4. We discuss inference and parameter estimation for LDA in Section 5. An illustrative example of fitting LDA to data is provided in Section 6. Empirical results in text modeling, text classification and collaborative filtering are presented in Section 7. Finally, Section 8 presents our conclusions.

2. Notation and terminology

We use the language of text collections throughout the paper, referring to entities such as "words," "documents," and "corpora." This is useful in that it helps to guide intuition, particularly when we introduce latent variables which aim to capture abstract notions such as topics. It is important to note, however, that the LDA model is not necessarily tied to text, and has applications to other problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics. Indeed, in Section 7.3, we present experimental results in the collaborative filtering domain.

Formally, we define the following terms:

- A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1,\ldots,V\}$. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.

- A *document* is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.

- A *corpus* is a collection of M documents denoted by $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

We wish to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other "similar" documents.

3. Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.[1]

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus $\mathcal{D}$:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n: → generate document
   (a) Choose a topic $z_n \boxed{\sim \text{Multinomial}(\theta).}$
   (b) Choose a word w_n from $p(w_n \,|\, z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1 \,|\, z^i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (θ and $\mathbf{z}$). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (a k-vector θ lies in the $(k-1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^k \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta \,|\, \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{1}$$

where the parameter α is a k-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient distribution on the simplex — it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 5, these properties will facilitate the development of inference and parameter estimation algorithms for LDA.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) = p(\theta \,|\, \alpha) \prod_{n=1}^{N} p(z_n \,|\, \theta) p(w_n \,|\, z_n, \beta), \tag{2}$$

[1] We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words.

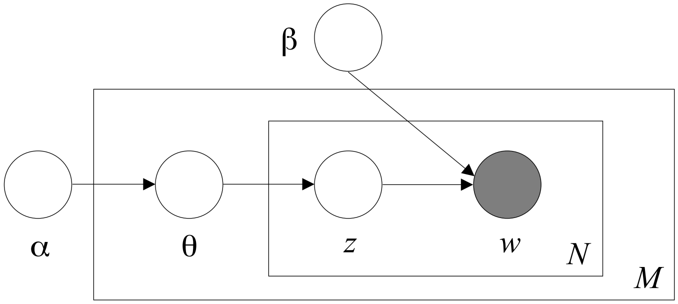


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

where $p(z_n \,|\, \theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w} \,|\, \alpha, \beta) = \int p(\theta \,|\, \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \,|\, \theta) p(w_n \,|\, z_n, \beta) \right) d\theta. \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(\mathcal{D} \,|\, \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \,|\, \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \,|\, \theta_d) p(w_{dn} \,|\, z_{dn}, \beta) \right) d\theta_d .$$

The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled *repeatedly* within the document. Under this model, documents can be associated with multiple topics.

Structures similar to that shown in Figure 1 are often studied in Bayesian statistical modeling, where they are referred to as *hierarchical models* (Gelman et al., 1995), or more precisely as *conditionally independent hierarchical models* (Kass and Steffey, 1989). Such models are also often referred to as *parametric empirical Bayes models*, a term that refers not only to a particular model structure, but also to the methods used for estimating parameters in the model (Morris, 1983). Indeed, as we discuss in Section 5, we adopt the empirical Bayes approach to estimating parameters such as α and β in simple implementations of LDA, but we also consider fuller Bayesian approaches as well.

3.1 LDA and exchangeability

A finite set of random variables $\{z_1, \ldots, z_N\}$ is said to be *exchangeable* if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(N)}).$$

An infinite sequence of random variables is *infinitely exchangeable* if every finite subsequence is exchangeable.

De Finetti's representation theorem states that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were *independent* and *identically distributed*, conditioned on that parameter.

In LDA, we assume that words are generated by topics (by fixed conditional distributions) and that those topics are infinitely exchangeable within a document. By de Finetti's theorem, the probability of a sequence of words and topics must therefore have the form:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \,|\, \theta) p(w_n \,|\, z_n) \right) d\theta,$$

where θ is the random parameter of a multinomial over topics. We obtain the LDA distribution on documents in Eq. (3) by marginalizing out the topic variables and endowing θ with a Dirichlet distribution.

3.2 A continuous mixture of unigrams

The LDA model shown in Figure 1 is somewhat more elaborate than the two-level models often studied in the classical hierarchical Bayesian literature. By marginalizing over the hidden topic variable z, however, we can understand LDA as a two-level model.

In particular, let us form the word distribution $p(w \,|\, \theta, \beta)$:

$$p(w \,|\, \theta, \beta) = \sum_{z} p(w \,|\, z, \beta) p(z \,|\, \theta).$$

Note that this is a random quantity since it depends on θ.

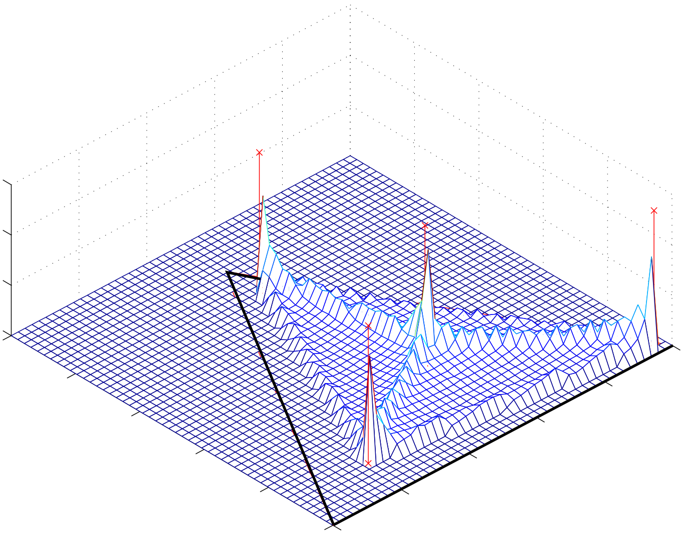


Figure 2: An example density on unigram distributions $p(w \mid \theta, \beta)$ under LDA for three words and four topics. The triangle embedded in the x-y plane is the 2-D simplex representing all possible multinomial distributions over three words. Each of the vertices of the triangle corresponds to a deterministic distribution that assigns probability one to one of the words; the midpoint of an edge gives probability 0.5 to two of the words; and the centroid of the triangle is the uniform distribution over all three words. The four points marked with an **x** are the locations of the multinomial distributions $p(w \mid z)$ for each of the four topics, and the surface shown on top of the simplex is an example of a density over the $(V-1)$-simplex (multinomial distributions of words) given by LDA.

We now define the following generative process for a document **w**:

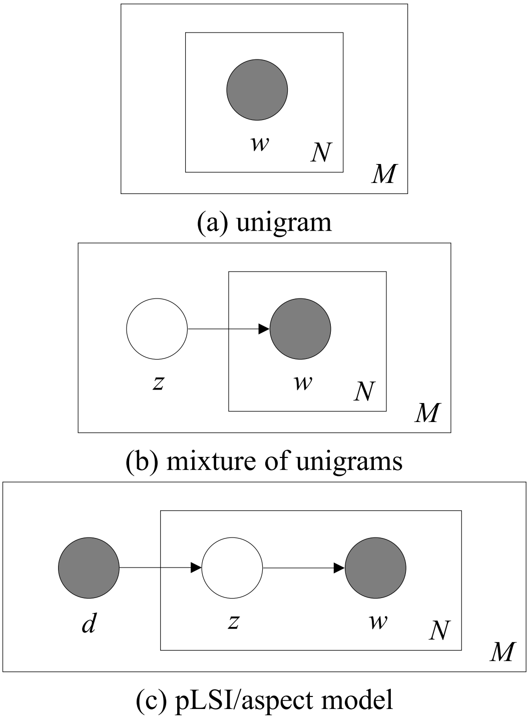
1. Choose $\theta \sim \mathrm{Dir}(\alpha)$.
2. For each of the N words w_n:
   (a) Choose a word w_n from $p(w_n \mid \theta, \beta)$.

This process defines the marginal distribution of a document as a continuous mixture distribution:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} p(w_n \mid \theta, \beta) \right) d\theta,$$

where $p(w_n \mid \theta, \beta)$ are the mixture components and $p(\theta \mid \alpha)$ are the mixture weights.

Figure 2 illustrates this interpretation of LDA. It depicts the distribution on $p(w \mid \theta, \beta)$ which is induced from a particular instance of an LDA model. Note that this distribution on the $(V-1)$-simplex is attained with only $k + kV$ parameters yet exhibits a very interesting multimodal structure.



(a) unigram

(b) mixture of unigrams

(c) pLSI/aspect model

Figure 3: Graphical model representation of different models of discrete data.

4. Relationship with other latent variable models

In this section we compare LDA to simpler latent variable models for text—the unigram model, a mixture of unigrams, and the pLSI model. Furthermore, we present a unified geometric interpretation of these models which highlights their key differences and similarities.

4.1 Unigram model

Under the unigram model, the words of every document are drawn independently from a single multinomial distribution:

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

This is illustrated in the graphical model in Figure 3a.

4.2 Mixture of unigrams

If we augment the unigram model with a discrete random topic variable z (Figure 3b), we obtain a *mixture of unigrams* model (Nigam et al., 2000). Under this mixture model, each document is generated by first choosing a topic z and then generating N words independently from the conditional multinomial $p(w \mid z)$. The probability of a document is:

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z).$$

When estimated from a corpus, the word distributions can be viewed as representations of topics under the assumption that each document exhibits exactly one topic. As the empirical results in Section 7 illustrate, this assumption is often too limiting to effectively model a large collection of documents.

In contrast, the LDA model allows documents to exhibit multiple topics to different degrees. This is achieved at a cost of just one additional parameter: there are $k-1$ parameters associated with $p(z)$ in the mixture of unigrams, versus the k parameters associated with $p(\theta\,|\,\alpha)$ in LDA.

4.3 Probabilistic latent semantic indexing

Probabilistic latent semantic indexing (pLSI) is another widely used document model (Hofmann, 1999). The pLSI model, illustrated in Figure 3c, posits that a document label d and a word w_n are conditionally independent given an unobserved topic z:

$$p(d,w_n) = p(d)\sum_{z} p(w_n\,|\,z)p(z\,|\,d).$$

The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic. In a sense, it does capture the possibility that a document may contain multiple topics since $p(z\,|\,d)$ serves as the mixture weights of the topics for a particular document d. However, it is important to note that d is a dummy index into the list of documents in the *training set*. Thus, d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures $p(z\,|\,d)$ only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of documents; there is no natural way to use it to assign probability to a previously unseen document.

A further difficulty with pLSI, which also stems from the use of a distribution indexed by training documents, is that the number of parameters which must be estimated grows linearly with the number of training documents. The parameters for a k-topic pLSI model are k multinomial distributions of size V and M mixtures over the k hidden topics. This gives $kV+kM$ parameters and therefore linear growth in M. The linear growth in parameters suggests that the model is prone to overfitting and, empirically, overfitting is indeed a serious problem (see Section 7.1). In practice, a tempering heuristic is used to smooth the parameters of the model for acceptable predictive performance. It has been shown, however, that overfitting can occur even when tempering is used (Popescul et al., 2001).

LDA overcomes both of these problems by treating the topic mixture weights as a k-parameter hidden *random variable* rather than a large set of individual parameters which are explicitly linked to the training set. As described in Section 3, LDA is a well-defined generative model and generalizes easily to new documents. Furthermore, the $k+kV$ parameters in a k-topic LDA model do not grow with the size of the training corpus. We will see in Section 7.1 that LDA does not suffer from the same overfitting issues as pLSI.

4.4 A geometric interpretation

A good way of illustrating the differences between LDA and the other latent topic models is by considering the geometry of the latent space, and seeing how a document is represented in that geometry under each model.

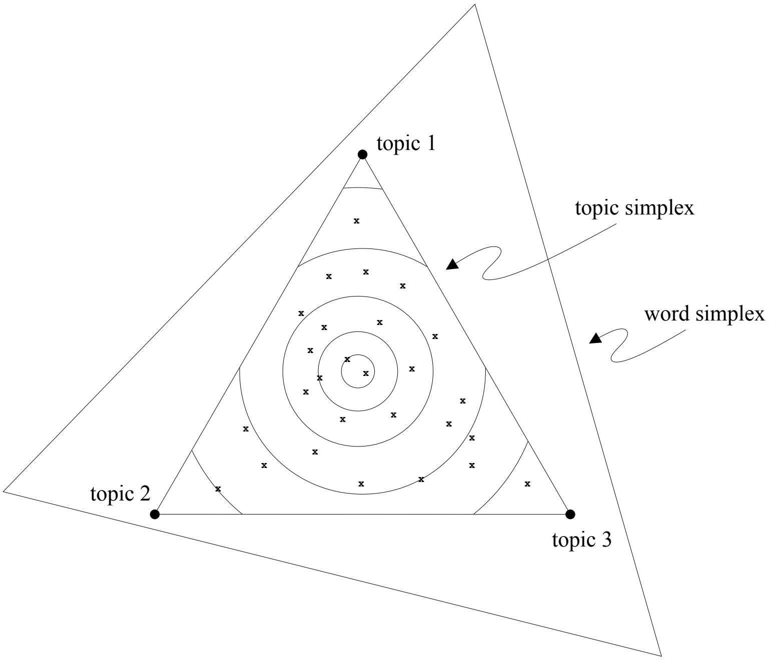


Figure 4: The topic simplex for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word (respectively) has probability one. The three points of the topic simplex correspond to three different distributions over words. The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by x. LDA places a smooth distribution on the topic simplex denoted by the contour lines.

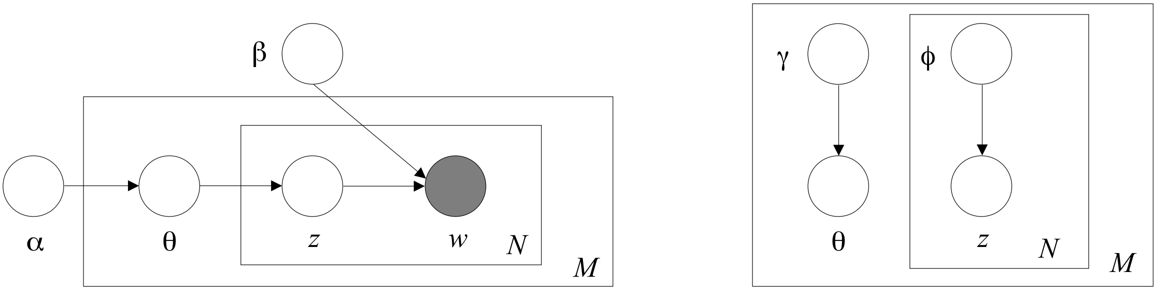


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

All four of the models described above—unigram, mixture of unigrams, pLSI, and LDA—operate in the space of distributions over words. Each such distribution can be viewed as a point on the $(V-1)$-simplex, which we call the word simplex.

The unigram model finds a single point on the word simplex and posits that all words in the corpus come from the corresponding distribution. The latent variable models consider k points on the word simplex and form a sub-simplex based on those points, which we call the topic simplex. Note that any point on the topic simplex is also a point on the word simplex. The different latent variable models use the topic simplex in different ways to generate a document.

- The mixture of unigrams model posits that for each document, one of the k points on the word simplex (that is, one of the corners of the topic simplex) is chosen randomly and all the words of the document are drawn from the distribution corresponding to that point.

- The pLSI model posits that each word of a *training* document comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics, i.e., a point on the topic simplex. There is one such distribution for each document; the set of training documents thus defines an empirical distribution on the topic simplex.

- LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter. This parameter is sampled once per document from a smooth distribution on the topic simplex.

These differences are highlighted in Figure 4.

5. Inference and Parameter Estimation

We have described the motivation behind LDA and illustrated its conceptual advantages over other latent topic models. In this section, we turn our attention to procedures for inference and parameter estimation under LDA.

5.1 Inference

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

$$p(\theta,\mathbf{z}\,|\,\mathbf{w},\alpha,\beta) = \frac{p(\theta,\mathbf{z},\mathbf{w}\,|\,\alpha,\beta)}{p(\mathbf{w}\,|\,\alpha,\beta)}.$$

Unfortunately, this distribution is intractable to compute in general. Indeed, to normalize the distribution we marginalize over the hidden variables and write Eq. (3) in terms of the model parameters:

$$p(\mathbf{w}\,|\,\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) d\theta,$$

a function which is intractable due to the coupling between θ and β in the summation over latent topics (Dickey, 1983). Dickey shows that this function is an expectation under a particular extension to the Dirichlet distribution which can be represented with special hypergeometric functions. It has been used in a Bayesian context for censored discrete data to represent the posterior on θ which, in that setting, is a random parameter (Dickey et al., 1987).

Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo (Jordan, 1999). In this section we describe a simple convexity-based variational algorithm for inference in LDA, and discuss some of the alternatives in Section 8.

5.2 Variational inference

The basic idea of convexity-based variational inference is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood (Jordan et al., 1999). Essentially, one considers a family of lower bounds, indexed by a set of *variational parameters*. The variational parameters are chosen by an optimization procedure that attempts to find the tightest possible lower bound.

A simple way to obtain a tractable family of lower bounds is to consider simple modifications of the original graphical model in which some of the edges and nodes are removed. Consider in particular the LDA model shown in Figure 5 (left). The problematic coupling between θ and β arises due to the edges between θ, $\mathbf{z}$, and $\mathbf{w}$. By dropping these edges and the $\mathbf{w}$ nodes, and endowing the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables. This family is characterized by the following variational distribution:

$$q(\theta,\mathbf{z}\,|\,\gamma,\phi) = q(\theta\,|\,\gamma) \prod_{n=1}^{N} q(z_n\,|\,\phi_n), \tag{4}$$

where the Dirichlet parameter γ and the multinomial parameters $(\phi_1,\ldots,\phi_N)$ are the free variational parameters.

Having specified a simplified family of probability distributions, the next step is to set up an optimization problem that determines the values of the variational parameters γ and ϕ. As we show in Appendix A, the desideratum of finding a tight lower bound on the log likelihood translates directly into the following optimization problem:

$$(\gamma^*,\phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta,\mathbf{z}\,|\,\gamma,\phi) \parallel p(\theta,\mathbf{z}\,|\,\mathbf{w},\alpha,\beta)). \tag{5}$$

```
initialize $\phi_{ni}^0 := 1/k$ for all $i$ and $n$
initialize $\gamma_i := \alpha_i + N/k$ for all $i$
**repeat**
    **for** $n = 1$ **to** $N$
        **for** $i = 1$ **to** $k$
            $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
        normalize $\phi_n^{t+1}$ to sum to 1.
    $\gamma^{t+1} := \alpha + \sum_{n=1}^N \phi_n^{t+1}$
**until** convergence
```

Figure 6: A variational inference algorithm for LDA.

Thus the optimizing values of the variational parameters are found by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior $p(\theta, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta)$. This minimization can be achieved via an iterative fixed-point method. In particular, we show in Appendix A.3 that by computing the derivatives of the KL divergence and setting them equal to zero, we obtain the following pair of update equations:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i) \,|\, \gamma]\} \tag{6}$$

$$\gamma_i = \alpha_i + \textstyle\sum_{n=1}^N \phi_{ni}. \tag{7}$$

As we show in Appendix A.1, the expectation in the multinomial update can be computed as follows:

$$\mathrm{E}_q[\log(\theta_i) \,|\, \gamma] = \Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right), \tag{8}$$

where Ψ is the first derivative of the $\log\Gamma$ function which is computable via Taylor approximations (Abramowitz and Stegun, 1970).

Eqs. (6) and (7) have an appealing intuitive interpretation. The Dirichlet update is a posterior Dirichlet given expected observations taken under the variational distribution, $\mathrm{E}[z_n \,|\, \phi_n]$. The multinomial update is akin to using Bayes' theorem, $p(z_n \,|\, w_n) \propto p(w_n \,|\, z_n) p(z_n)$, where $p(z_n)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution.

It is important to note that the variational distribution is actually a conditional distribution, varying as a function of $\mathbf{w}$. This occurs because the optimization problem in Eq. (5) is conducted for fixed $\mathbf{w}$, and thus yields optimizing parameters (γ^*, ϕ^*) that are a function of $\mathbf{w}$. We can write the resulting variational distribution as $q(\theta, \mathbf{z} \,|\, \gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$, where we have made the dependence on $\mathbf{w}$ explicit. Thus the variational distribution can be viewed as an approximation to the posterior distribution $p(\theta, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta)$.

In the language of text, the optimizing parameters $(\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$ are document-specific. In particular, we view the Dirichlet parameters $\gamma^*(\mathbf{w})$ as providing a representation of a document in the topic simplex.

We summarize the variational inference procedure in Figure 6, with appropriate starting points for γ and ϕ_n. From the pseudocode it is clear that each iteration of variational inference for LDA requires $\mathrm{O}((N+1)k)$ operations. Empirically, we find that the number of iterations required for a single document is on the order of the number of words in the document. This yields a total number of operations roughly on the order of N^2k.

docs ↔ topics ; E step memberships(EM)
topics ↔ words : M step param est (EM)

5.3 Parameter estimation

In this section we present an empirical Bayes method for parameter estimation in the LDA model (see Section 5.4 for a fuller Bayesian approach). In particular, given a corpus of documents $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$, we wish to find parameters α and β that maximize the (marginal) log likelihood of the data:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \,|\, \alpha, \beta).$$

As we have described above, the quantity $p(\mathbf{w} \,|\, \alpha, \beta)$ cannot be computed tractably. However, variational inference provides us with a tractable lower bound on the log likelihood, a bound which we can maximize with respect to α and β. We can thus find approximate empirical Bayes estimates for the LDA model via an alternating *variational EM* procedure that maximizes a lower bound with respect to the variational parameters γ and ϕ, and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters α and β.

We provide a detailed derivation of the variational EM algorithm for LDA in Appendix A.4. The derivation yields the following iterative algorithm:

1. (E-step) For each document, find the optimizing values of the variational parameters $\{\gamma_d^*, \phi_d^* : d \in \mathcal{D}\}$. This is done as described in the previous section.

2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β. This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step.

These two steps are repeated until the lower bound on the log likelihood converges.

In Appendix A.4, we show that the M-step update for the conditional multinomial parameter β can be written out analytically:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j. \tag{9}$$

We further show that the M-step update for Dirichlet parameter α can be implemented using an efficient Newton-Raphson method in which the Hessian is inverted in linear time.

5.4 Smoothing

The large vocabulary size that is characteristic of many document corpora creates serious problems of sparsity. A new document is very likely to contain words that did not appear in any of the documents in a training corpus. Maximum likelihood estimates of the multinomial parameters assign zero probability to such words, and thus zero probability to new documents. The standard approach to coping with this problem is to "smooth" the multinomial parameters, assigning positive probability to all vocabulary items whether or not they are observed in the training set (Jelinek, 1997). Laplace smoothing is commonly used; this essentially yields the mean of the posterior distribution under a uniform Dirichlet prior on the multinomial parameters.

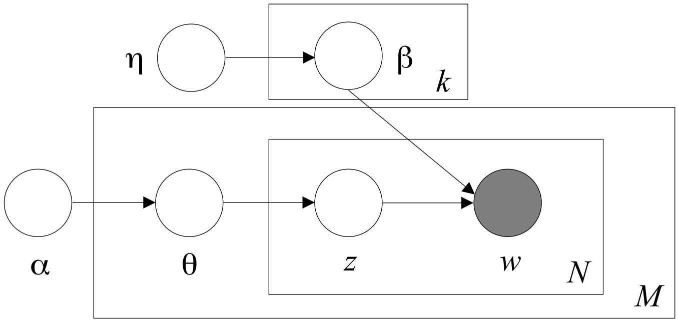


Figure 7: Graphical model representation of the smoothed LDA model.

Unfortunately, in the mixture model setting, simple Laplace smoothing is no longer justified as a maximum a posteriori method (although it is often implemented in practice; cf. Nigam et al., 1999). In fact, by placing a Dirichlet prior on the multinomial parameter we obtain an intractable posterior in the mixture model setting, for much the same reason that one obtains an intractable posterior in the basic LDA model. Our proposed solution to this problem is to simply apply variational inference methods to the extended model that includes Dirichlet smoothing on the multinomial parameter.

In the LDA setting, we obtain the extended graphical model shown in Figure 7. We treat β as a $k \times V$ random matrix (one row for each mixture component), where we assume that each row is independently drawn from an exchangeable Dirichlet distribution.[2] We now extend our inference procedures to treat the β_i as random variables that are endowed with a posterior distribution, conditioned on the data. Thus we move beyond the empirical Bayes procedure of Section 5.3 and consider a fuller Bayesian approach to LDA.

We consider a variational approach to Bayesian inference that places a separable distribution on the random variables β, θ, and $\mathbf{z}$ (Attias, 2000):

$$q(\beta_{1:k}, \mathbf{z}_{1:M}, \theta_{1:M} \,|\, \lambda, \phi, \gamma) = \prod_{i=1}^{k} \mathrm{Dir}(\beta_i \,|\, \lambda_i) \prod_{d=1}^{M} q_d(\theta_d, \mathbf{z}_d \,|\, \phi_d, \gamma_d),$$

where $q_d(\theta, \mathbf{z} \,|\, \phi, \gamma)$ is the variational distribution defined for LDA in Eq. (4). As is easily verified, the resulting variational inference procedure again yields Eqs. (6) and (7) as the update equations for the variational parameters ϕ and γ, respectively, as well as an additional update for the new variational parameter λ:

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}.$$

Iterating these equations to convergence yields an approximate posterior distribution on β, θ, and $\mathbf{z}$.

We are now left with the hyperparameter η on the exchangeable Dirichlet, as well as the hyperparameter α from before. Our approach to setting these hyperparameters is again (approximate) empirical Bayes—we use variational EM to find maximum likelihood estimates of these parameters based on the marginal likelihood. These procedures are described in Appendix A.4.

[2] An exchangeable Dirichlet is simply a Dirichlet distribution with a single scalar parameter η. The density is the same as a Dirichlet (Eq. 1) where $\alpha_i = \eta$ for each component.

6. Example

In this section, we provide an illustrative example of the use of an LDA model on real data. Our data are 16,000 documents from a subset of the TREC AP corpus (Harman, 1992). After removing a standard list of stop words, we used the EM algorithm described in Section 5.3 to find the Dirichlet and conditional multinomial parameters for a 100-topic LDA model. The top words from some of the resulting multinomial distributions $p(w \mid z)$ are illustrated in Figure 8 (top). As we have hoped, these distributions seem to capture some of the underlying topics in the corpus (and we have named them according to these topics).

As we emphasized in Section 4, one of the advantages of LDA over related latent variable models is that it provides well-defined inference procedures for previously unseen documents. Indeed, we can illustrate how LDA works by performing inference on a held-out document and examining the resulting variational posterior parameters.

Figure 8 (bottom) is a document from the TREC AP corpus which was not used for parameter estimation. Using the algorithm in Section 5.1, we computed the variational posterior Dirichlet parameters γ for the article and variational posterior multinomial parameters ϕ_n for each word in the article.

Recall that the ith posterior Dirichlet parameter γ_i is approximately the ith prior Dirichlet parameter α_i plus the expected number of words which were generated by the ith topic (see Eq. 7). Therefore, the prior Dirichlet parameters subtracted from the posterior Dirichlet parameters indicate the expected number of words which were allocated to each topic for a particular document. For the example article in Figure 8 (bottom), most of the γ_i are close to α_i. Four topics, however, are significantly larger (by this, we mean $\gamma_i - \alpha_i \geq 1$). Looking at the corresponding distributions over words identifies the topics which mixed to form this document (Figure 8, top).

Further insight comes from examining the ϕ_n parameters. These distributions approximate $p(z_n \mid \mathbf{w})$ and tend to peak towards one of the k possible topic values. In the article text in Figure 8, the words are color coded according to these values (i.e., the ith color is used if $q_n(z_n^i = 1) > 0.9$). With this illustration, one can identify how the different topics mixed in the document text.

While demonstrating the power of LDA, the posterior analysis also highlights some of its limitations. In particular, the bag-of-words assumption allows words that should be generated by the same topic (e.g., "William Randolph Hearst Foundation") to be allocated to several different topics. Overcoming this limitation would require some form of extension of the basic LDA model; in particular, we might relax the bag-of-words assumption by assuming partial exchangeability or Markovianity of word sequences.

7. Applications and Empirical Results

In this section, we discuss our empirical evaluation of LDA in several problem domains—document modeling, document classification, and collaborative filtering.

In all of the mixture models, the expected complete log likelihood of the data has local maxima at the points where all or some of the mixture components are equal to each other. To avoid these local maxima, it is important to initialize the EM algorithm appropriately. In our experiments, we initialize EM by seeding each conditional multinomial distribution with five documents, reducing their effective total length to two words, and smoothing across the whole vocabulary. This is essentially an approximation to the scheme described in Heckerman and Meila (2001).

“Arts”	**“Budgets”**	**“Children”**	**“Education”**
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Num. topics (k)	Perplexity (Mult. Mixt.)	Perplexity (pLSI)
2	22,266	7,052
5	2.20×10^{8}	17,588
10	1.93×10^{17}	63,800
20	1.20×10^{22}	2.52×10^{5}
50	4.19×10^{106}	5.04×10^{6}
100	2.39×10^{150}	1.72×10^{7}
200	3.51×10^{264}	1.31×10^{7}

Table 1: Overfitting in the mixture of unigrams and pLSI models for the AP corpus. Similar behavior is observed in the nematode corpus (not reported).

7.1 Document modeling

We trained a number of latent variable models, including LDA, on two text corpora to compare the generalization performance of these models. The documents in the corpora are treated as unlabeled; thus, our goal is density estimation—we wish to achieve high likelihood on a held-out test set. In particular, we computed the *perplexity* of a held-out test set to evaluate the models. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraicly equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance.[3] More formally, for a test set of M documents, the perplexity is:

$$perplexity(\mathcal{D}_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

In our experiments, we used a corpus of scientific abstracts from the C. Elegans community (Avery, 2002) containing 5,225 abstracts with 28,414 unique terms, and a subset of the TREC AP corpus containing 16,333 newswire articles with 23,075 unique terms. In both cases, we held out 10% of the data for test purposes and trained the models on the remaining 90%. In preprocessing the data, we removed a standard list of 50 stop words from each corpus. From the AP data, we further removed words that occurred only once.

We compared LDA with the unigram, mixture of unigrams, and pLSI models described in Section 4. We trained all the hidden variable models using EM with exactly the same stopping criteria, that the average change in expected log likelihood is less than 0.001%.

Both the pLSI model and the mixture of unigrams suffer from serious overfitting issues, though for different reasons. This phenomenon is illustrated in Table 1. In the mixture of unigrams model, overfitting is a result of peaked posteriors in the training set; a phenomenon familiar in the supervised setting, where this model is known as the naive Bayes model (Rennie, 2001). This leads to a

[3] Note that we simply use perplexity as a figure of merit for comparing models. The models that we compare are all unigram ("bag-of-words") models, which—as we have discussed in the Introduction—are of interest in the information retrieval context. We are *not* attempting to do language modeling in this paper—an enterprise that would require us to examine trigram or other higher-order models. We note in passing, however, that extensions of LDA could be considered that involve Dirichlet-multinomial over trigrams instead of unigrams. We leave the exploration of such extensions to language modeling to future work.

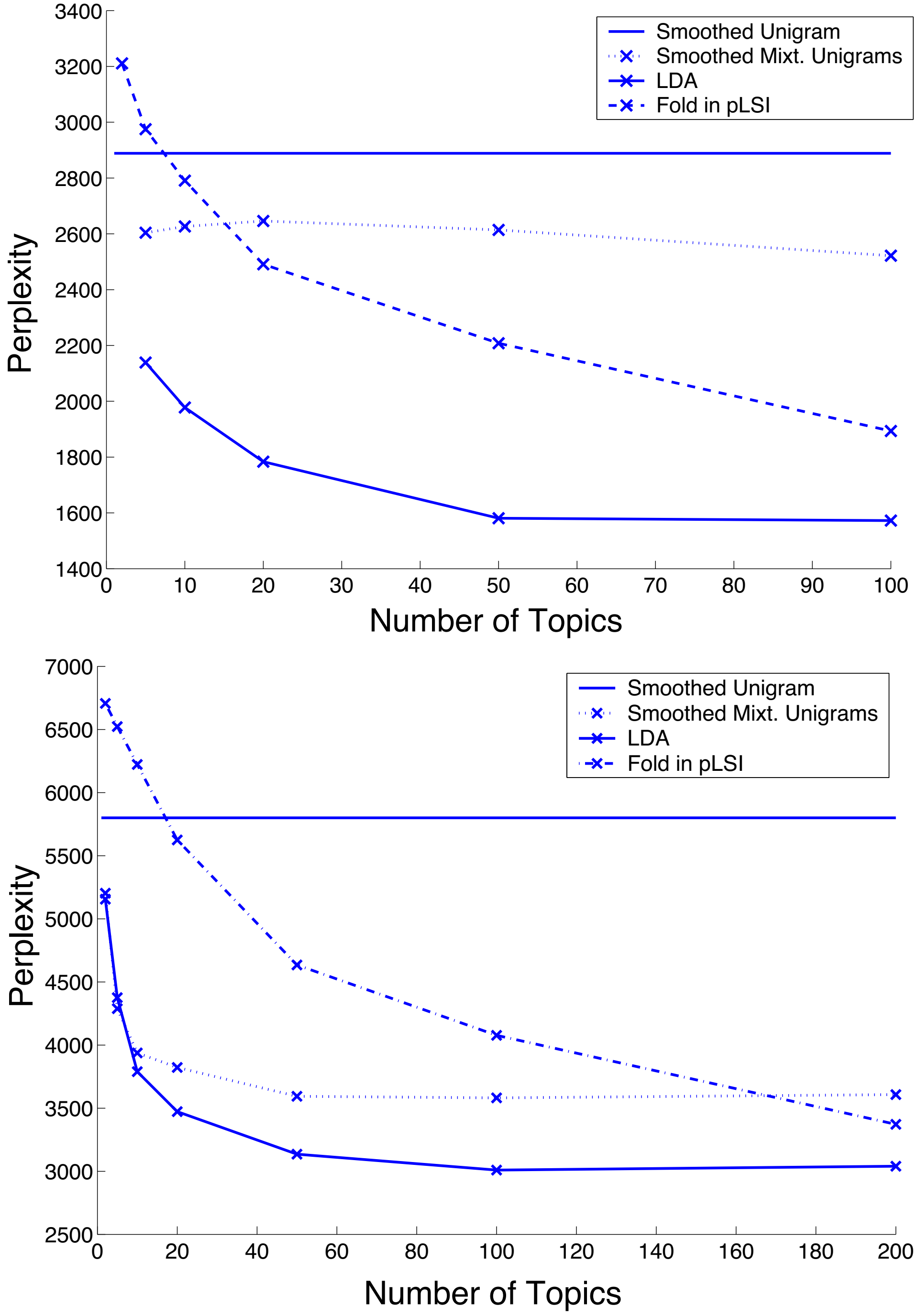


Figure 9: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.

nearly deterministic clustering of the training documents (in the E-step) which is used to determine the word probabilities in each mixture component (in the M-step). A previously unseen document may best fit one of the resulting mixture components, but will probably contain at least one word which did not occur in the training documents that were assigned to that component. Such words will have a very small probability, which causes the perplexity of the new document to explode. As k increases, the documents of the training corpus are partitioned into finer collections and thus induce more words with small probabilities.

In the mixture of unigrams, we can alleviate overfitting through the variational Bayesian smoothing scheme presented in Section 5.4. This ensures that all words will have some probability under every mixture component.

In the pLSI case, the hard clustering problem is alleviated by the fact that each document is allowed to exhibit a different proportion of topics. However, pLSI only refers to the training documents and a different overfitting problem arises that is due to the dimensionality of the $p(z|d)$ parameter. One reasonable approach to assigning probability to a previously unseen document is by marginalizing over d:

$$p(\mathbf{w}) = \sum_{d} \prod_{n=1}^{N} \sum_{z} p(w_n \,|\, z) p(z \,|\, d) p(d).$$

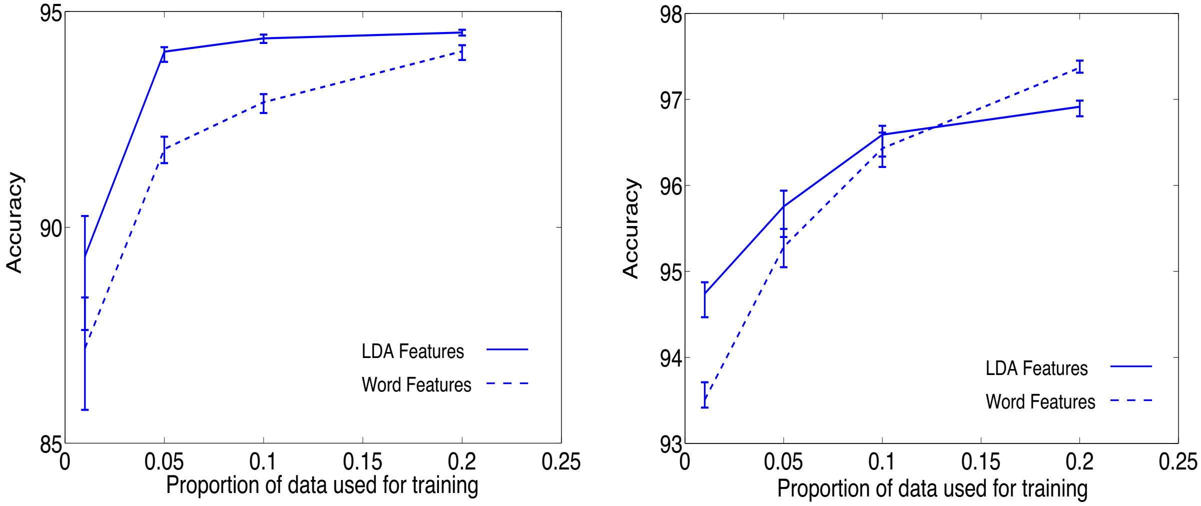
Essentially, we are integrating over the empirical distribution on the topic simplex (see Figure 4).

This method of inference, though theoretically sound, causes the model to overfit. The document-specific topic distribution has some components which are close to zero for those topics that do not appear in the document. Thus, certain words will have very small probability in the estimates of each mixture component. When determining the probability of a new document through marginalization, only those training documents which exhibit a similar proportion of topics will contribute to the likelihood. For a given training document's topic proportions, any word which has small probability in all the constituent topics will cause the perplexity to explode. As k gets larger, the chance that a training document will exhibit topics that cover all the words in the new document decreases and thus the perplexity grows. Note that pLSI does not overfit as quickly (with respect to k) as the mixture of unigrams.

This overfitting problem essentially stems from the restriction that each future document exhibit the same topic proportions as were seen in one or more of the training documents. Given this constraint, we are not free to choose the most likely proportions of topics for the new document. An alternative approach is the "folding-in" heuristic suggested by Hofmann (1999), where one ignores the $p(z|d)$ parameters and refits $p(z|d_{\text{new}})$. Note that this gives the pLSI model an unfair advantage by allowing it to refit $k-1$ parameters to the test data.

LDA suffers from neither of these problems. As in pLSI, each document can exhibit a different proportion of underlying topics. However, LDA can easily assign probability to a new document; no heuristics are needed for a new document to be endowed with a different set of topic proportions than were associated with documents in the training corpus.

Figure 9 presents the perplexity for each model on both corpora for different values of k. The pLSI model and mixture of unigrams are suitably corrected for overfitting. The latent variable models perform better than the simple unigram model. LDA consistently performs better than the other models.



(a)

(b)

Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

7.2 Document classification

In the text classification problem, we wish to classify a document into two or more mutually exclusive classes. As in any classification problem, we may wish to consider generative approaches or discriminative approaches. In particular, by using one LDA module for each class, we obtain a generative model for classification. It is also of interest to use LDA in the discriminative framework, and this is our focus in this section.

A challenging aspect of the document classification problem is the choice of features. Treating individual words as features yields a rich but very large feature set (Joachims, 1999). One way to reduce this feature set is to use an LDA model for dimensionality reduction. In particular, LDA reduces any document to a fixed set of real-valued features—the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ associated with the document. It is of interest to see how much discriminatory information we lose in reducing the document description to these parameters.

We conducted two binary classification experiments using the Reuters-21578 dataset. The dataset contains 8000 documents and 15,818 words.

In these experiments, we estimated the parameters of an LDA model on all the documents, without reference to their true class label. We then trained a support vector machine (SVM) on the low-dimensional representations provided by LDA and compared this SVM to an SVM trained on all the word features.

Using the SVMLight software package (Joachims, 1999), we compared an SVM trained on all the word features with those trained on features induced by a 50-topic LDA model. Note that we reduce the feature space by 99.6 percent in this case.

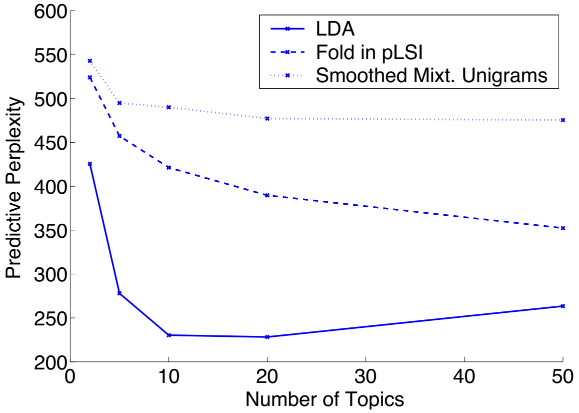


Figure 11: Results for collaborative filtering on the EachMovie data.

Figure 10 shows our results. We see that there is little reduction in classification performance in using the LDA-based features; indeed, in almost all cases the performance is improved with the LDA features. Although these results need further substantiation, they suggest that the topic-based representation provided by LDA may be useful as a fast filtering algorithm for feature selection in text classification.

7.3 Collaborative filtering

Our final experiment uses the EachMovie collaborative filtering data. In this data set, a collection of users indicates their preferred movie choices. A user and the movies chosen are analogous to a document and the words in the document (respectively).

The collaborative filtering task is as follows. We train a model on a fully observed set of users. Then, for each unobserved user, we are shown all but one of the movies preferred by that user and are asked to predict what the held-out movie is. The different algorithms are evaluated according to the likelihood they assign to the held-out movie. More precisely, define the predictive perplexity on M test users to be:

$$\textit{predictive-perplexity}(\mathcal{D}_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_{d,N_d} \,|\, \mathbf{w}_{d,1:N_d-1})}{M})\right\}.$$

We restricted the EachMovie dataset to users that positively rated at least 100 movies (a positive rating is at least four out of five stars). We divided this set of users into 3300 training users and 390 testing users.

Under the mixture of unigrams model, the probability of a movie given a set of observed movies is obtained from the posterior distribution over topics:

$$p(w|\mathbf{w}_{\text{obs}}) = \sum_{z} p(w|z)p(z|\mathbf{w}_{\text{obs}}).$$

In the pLSI model, the probability of a held-out movie is given by the same equation except that $p(z|\mathbf{w}_{\text{obs}})$ is computed by folding in the previously seen movies. Finally, in the LDA model, the probability of a held-out movie is given by integrating over the posterior Dirichlet:

$$p(w|\mathbf{w}_{\text{obs}}) = \int \sum_{z} p(w|z)p(z|\theta)p(\theta|\mathbf{w}_{\text{obs}})d\theta,$$

where $p(\theta|\mathbf{w}_{\text{obs}})$ is given by the variational inference method described in Section 5.2. Note that this quantity is efficient to compute. We can interchange the sum and integral sign, and compute a linear combination of k Dirichlet expectations.

With a vocabulary of 1600 movies, we find the predictive perplexities illustrated in Figure 11. Again, the mixture of unigrams model and pLSI are corrected for overfitting, but the best predictive perplexities are obtained by the LDA model.

8. Discussion

We have described latent Dirichlet allocation, a flexible generative probabilistic model for collections of discrete data. LDA is based on a simple exchangeability assumption for the words and topics in a document; it is therefore realized by a straightforward application of de Finetti's representation theorem. We can view LDA as a dimensionality reduction technique, in the spirit of LSI, but with proper underlying generative probabilistic semantics that make sense for the type of data that it models.

Exact inference is intractable for LDA, but any of a large suite of approximate inference algorithms can be used for inference and parameter estimation within the LDA framework. We have presented a simple convexity-based variational approach for inference, showing that it yields a fast algorithm resulting in reasonable comparative performance in terms of test set likelihood. Other approaches that might be considered include Laplace approximation, higher-order variational techniques, and Monte Carlo methods. In particular, Leisink and Kappen (2002) have presented a general methodology for converting low-order variational lower bounds into higher-order variational bounds. It is also possible to achieve higher accuracy by dispensing with the requirement of maintaining a bound, and indeed Minka and Lafferty (2002) have shown that improved inferential accuracy can be obtained for the LDA model via a higher-order variational technique known as expectation propagation. Finally, Griffiths and Steyvers (2002) have presented a Markov chain Monte Carlo algorithm for LDA.

LDA is a simple model, and although we view it as a competitor to methods such as LSI and pLSI in the setting of dimensionality reduction for document collections and other discrete corpora, it is also intended to be illustrative of the way in which probabilistic models can be scaled up to provide useful inferential machinery in domains involving multiple levels of structure. Indeed, the principal advantages of generative models such as LDA include their modularity and their extensibility. As a probabilistic module, LDA can be readily embedded in a more complex model—a property that is not possessed by LSI. In recent work we have used pairs of LDA modules to model relationships between images and their corresponding descriptive captions (Blei and Jordan, 2002). Moreover, there are numerous possible extensions of LDA. For example, LDA is readily extended to continuous data or other non-multinomial data. As is the case for other mixture models, including finite mixture models and hidden Markov models, the "emission" probability $p(w_n|z_n)$ contributes only a likelihood value to the inference procedures for LDA, and other likelihoods are readily substituted in its place. In particular, it is straightforward to develop a continuous variant of LDA in which Gaussian observables are used in place of multinomials. Another simple extension of LDA comes from allowing mixtures of Dirichlet distributions in the place of the single Dirichlet of LDA. This allows a richer structure in the latent topic space and in particular allows a form of document clustering that is different from the clustering that is achieved via shared topics. Finally, a variety of extensions of LDA can be considered in which the distributions on the topic variables are elaborated. For example, we could arrange the topics in a time series, essentially relaxing the full exchangeability assumption to one of partial exchangeability. We could also consider partially exchangeable models in which we condition on exogenous variables; thus, for example, the topic distribution could be conditioned on features such as "paragraph" or "sentence," providing a more powerful text model that makes use of information obtained from a parser.

Acknowledgements

This work was supported by the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637). Andrew Y. Ng and David M. Blei were additionally supported by fellowships from the Microsoft Corporation.

References

M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1970.

D. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin, 1985.

H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, 2000.

L. Avery. Caenorrhabditis genetic center bibliography. 2002. URL `http://elegans.swmed.edu/wli/cgcbib`.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

D. Blei and M. Jordan. Modeling annotated data. Technical Report UCB//CSD-02-1202, U.C. Berkeley Computer Science Division, 2002.

B. de Finetti. *Theory of probability. Vol. 1-2*. John Wiley & Sons Ltd., Chichester, 1990. Reprint of the 1975 translation.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

P. Diaconis. Recent progress on de Finetti's notions of exchangeability. In *Bayesian statistics, 3 (Valencia, 1987)*, pages 111–125. Oxford Univ. Press, New York, 1988.

J. Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78:628–637, 1983.

J. Dickey, J. Jiang, and J. Kadane. Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, 82:773–781, 1987.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman & Hall, London, 1995.

T. Griffiths and M. Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.

D. Harman. Overview of the first text retrieval conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 1–20, 1992.

D. Heckerman and M. Meila. An experimental comparison of several clustering and initialization methods. *Machine Learning*, 42:9–29, 2001.

T. Hofmann. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1997.

T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. M.I.T. Press, 1999.

M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

R. Kass and D. Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84 (407):717–726, 1989.

M. Leisink and H. Kappen. General lower bounds based on computer generated higher order expansions. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 2002.

T. Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.

T. P. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence (UAI)*, 2002.

C. Morris. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–65, 1983. With discussion.

K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

C. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. pages 159–168, 1998.

A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*, 2001.

J. Rennie. Improving multi-class text classification with naive Bayes. Technical Report AITR-2001-004, M.I.T., 2001.

G. Ronning. Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistcal Computation and Simulation*, 34(4):215–221, 1989.

G. Salton and M. McGill, editors. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

Appendix A. Inference and parameter estimation

In this appendix, we derive the variational inference procedure (Eqs. 6 and 7) and the parameter maximization procedure for the conditional multinomial (Eq. 9) and for the Dirichlet. We begin by deriving a useful property of the Dirichlet distribution.

A.1 Computing $\mathbf{E}[\log(\theta_i \,|\, \alpha)]$

The need to compute the expected value of the log of a single probability component under the Dirichlet arises repeatedly in deriving the inference and parameter estimation procedures for LDA. This value can be easily computed from the natural parameterization of the exponential family representation of the Dirichlet distribution.

Recall that a distribution is in the exponential family if it can be written in the form:

$$p(x \,|\, \eta) = h(x) \exp\left\{\eta^T T(x) - A(\eta)\right\},$$

where η is the natural parameter, $T(x)$ is the sufficient statistic, and $A(\eta)$ is the log of the normalization factor.

We can write the Dirichlet in this form by exponentiating the log of Eq. (1):

$$p(\theta \,|\, \alpha) = \exp\left\{\left(\textstyle\sum_{i=1}^k (\alpha_i - 1)\log\theta_i\right) + \log\Gamma\left(\textstyle\sum_{i=1}^k \alpha_i\right) - \textstyle\sum_{i=1}^k \log\Gamma(\alpha_i)\right\}.$$

From this form, we immediately see that the natural parameter of the Dirichlet is $\eta_i = \alpha_i - 1$ and the sufficient statistic is $T(\theta_i) = \log\theta_i$. Furthermore, using the general fact that the derivative of the log normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic, we obtain:

$$\mathrm{E}[\log\theta_i \,|\, \alpha] = \Psi(\alpha_i) - \Psi\left(\textstyle\sum_{j=1}^k \alpha_j\right)$$

where Ψ is the digamma function, the first derivative of the log Gamma function.

A.2 Newton-Raphson methods for a Hessian with special structure

In this section we describe a linear algorithm for the usually cubic Newton-Raphson optimization method. This method is used for maximum likelihood estimation of the Dirichlet distribution (Ronning, 1989, Minka, 2000).

The Newton-Raphson optimization technique finds a stationary point of a function by iterating:

$$\alpha_{\text{new}} = \alpha_{\text{old}} - H(\alpha_{\text{old}})^{-1} g(\alpha_{\text{old}})$$

where $H(\alpha)$ and $g(\alpha)$ are the Hessian matrix and gradient respectively at the point α. In general, this algorithm scales as $\mathrm{O}(N^3)$ due to the matrix inversion.

If the Hessian matrix is of the form:

$$H = \mathrm{diag}(h) + \mathbf{1} z \mathbf{1}^{\mathrm{T}}, \tag{10}$$

where $\mathrm{diag}(h)$ is defined to be a diagonal matrix with the elements of the vector h along the diagonal, then we can apply the matrix inversion lemma and obtain:

$$H^{-1} = \mathrm{diag}(h)^{-1} - \frac{\mathrm{diag}(h)^{-1} \mathbf{1}\mathbf{1}^{\mathrm{T}} \mathrm{diag}(h)^{-1}}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}$$

Multiplying by the gradient, we obtain the ith component:

$$(H^{-1} g)_i = \frac{g_i - c}{h_i}$$

where

$$c = \frac{\sum_{j=1}^{k} g_j / h_j}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}.$$

Observe that this expression depends only on the $2k$ values h_i and g_i and thus yields a Newton-Raphson algorithm that has linear time complexity.

A.3 Variational inference

In this section we derive the variational inference algorithm described in Section 5.1. Recall that this involves using the following *variational distribution*:

$$q(\theta, \mathbf{z} \,|\, \gamma, \phi) = q(\theta \,|\, \gamma) \prod_{n=1}^{N} q(z_n \,|\, \phi_n) \tag{11}$$

as a surrogate for the posterior distribution $p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta)$, where the *variational parameters* γ and ϕ are set via an optimization procedure that we now describe.

Following Jordan et al. (1999), we begin by bounding the log likelihood of a document using Jensen's inequality. Omitting the parameters γ and ϕ for simplicity, we have:

$$\begin{aligned}
\log p(\mathbf{w} \,|\, \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) d\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\
&\geq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta) d\theta - \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log q(\theta, \mathbf{z}) d\theta \\
&= \mathrm{E}_q[\log p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta)] - \mathrm{E}_q[\log q(\theta, \mathbf{z})].
\end{aligned} \tag{12}$$

Thus we see that Jensen's inequality provides us with a lower bound on the log likelihood for an arbitrary variational distribution $q(\theta, \mathbf{z} \mid \gamma, \phi)$.

It can be easily verified that the difference between the left-hand side and the right-hand side of the Eq. (12) is the KL divergence between the variational posterior probability and the true posterior probability. That is, letting $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ denote the right-hand side of Eq. (12) (where we have restored the dependence on the variational parameters γ and ϕ in our notation), we have:

$$\log p(\mathbf{w} \mid \alpha, \beta) = \mathcal{L}(\gamma, \phi; \alpha, \beta) + \mathrm{D}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)). \tag{13}$$

This shows that maximizing the lower bound $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ with respect to γ and ϕ is equivalent to minimizing the KL divergence between the variational posterior probability and the true posterior probability, the optimization problem presented earlier in Eq. (5).

We now expand the lower bound by using the factorizations of p and q:

$$\begin{aligned} \mathcal{L}(\gamma, \phi; \alpha, \beta) = {} & \mathrm{E}_q[\log p(\theta \mid \alpha)] + \mathrm{E}_q[\log p(\mathbf{z} \mid \theta)] + \mathrm{E}_q[\log p(\mathbf{w} \mid \mathbf{z}, \beta)] \\ & - \mathrm{E}_q[\log q(\theta)] - \mathrm{E}_q[\log q(\mathbf{z})]. \end{aligned} \tag{14}$$

Finally, we expand Eq. (14) in terms of the model parameters (α, β) and the variational parameters (γ, ϕ). Each of the five lines below expands one of the five terms in the bound:

$$\begin{aligned} \mathcal{L}(\gamma, \phi; \alpha, \beta) = {} & \log \Gamma\left(\textstyle\sum_{j=1}^{k} \alpha_j\right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ & + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ & + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij} \\ & - \log \Gamma\left(\textstyle\sum_{j=1}^{k} \gamma_j\right) + \sum_{i=1}^{k} \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \\ & - \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}, \end{aligned} \tag{15}$$

where we have made use of Eq. (8).

In the following two sections, we show how to maximize this lower bound with respect to the variational parameters ϕ and γ.

A.3.1 VARIATIONAL MULTINOMIAL

We first maximize Eq. (15) with respect to ϕ_{ni}, the probability that the nth word is generated by latent topic i. Observe that this is a constrained maximization since $\sum_{i=1}^{k} \phi_{ni} = 1$.

We form the Lagrangian by isolating the terms which contain ϕ_{ni} and adding the appropriate Lagrange multipliers. Let β_{iv} be $p(w_n^v = 1 \mid z^i = 1)$ for the appropriate v. (Recall that each w_n is a vector of size V with exactly one component equal to one; we can select the unique v such that $w_n^v = 1$):

$$\mathcal{L}_{[\phi_{ni}]} = \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) + \phi_{ni} \log \beta_{iv} - \phi_{ni} \log \phi_{ni} + \lambda_n \left(\textstyle\sum_{j=1}^{k} \phi_{ni} - 1\right),$$

where we have dropped the arguments of $\mathcal{L}$ for simplicity, and where the subscript ϕ_{ni} denotes that we have retained only those terms in $\mathcal{L}$ that are a function of ϕ_{ni}. Taking derivatives with respect to ϕ_{ni}, we obtain:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda.$$

Setting this derivative to zero yields the maximizing value of the variational parameter ϕ_{ni} (cf. Eq. 6):

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right). \tag{16}$$

A.3.2 VARIATIONAL DIRICHLET

Next, we maximize Eq. (15) with respect to γ_i, the ith component of the posterior Dirichlet parameter. The terms containing γ_i are:

$$\begin{aligned}\mathcal{L}_{[\gamma]} = {} & \sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) + \sum_{n=1}^{N} \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) \\ & - \log \Gamma\left(\sum_{j=1}^{k} \gamma_j\right) + \log \Gamma(\gamma_i) - \sum_{i=1}^{k} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right).\end{aligned}$$

This simplifies to:

$$\mathcal{L}_{[\gamma]} = \sum_{i=1}^{k} \left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right)\left(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i\right) - \log \Gamma\left(\sum_{j=1}^{k} \gamma_j\right) + \log \Gamma(\gamma_i).$$

We take the derivative with respect to γ_i:

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i\right) - \Psi'\left(\sum_{j=1}^{k} \gamma_j\right) \sum_{j=1}^{k} \left(\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j\right).$$

Setting this equation to zero yields a maximum at:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}. \tag{17}$$

Since Eq. (17) depends on the variational multinomial ϕ, full variational inference requires alternating between Eqs. (16) and (17) until the bound converges.

A.4 Parameter estimation

In this final section, we consider the problem of obtaining empirical Bayes estimates of the model parameters α and β. We solve this problem by using the variational lower bound as a surrogate for the (intractable) marginal log likelihood, with the variational parameters ϕ and γ fixed to the values found by variational inference. We then obtain (approximate) empirical Bayes estimates by maximizing this lower bound with respect to the model parameters.

We have thus far considered the log likelihood for a single document. Given our assumption of exchangeability for the documents, the overall log likelihood of a corpus $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ is the sum of the log likelihoods for individual documents; moreover, the overall variational lower bound is the sum of the individual variational bounds. In the remainder of this section, we abuse notation by using $\mathcal{L}$ for the total variational bound, indexing the document-specific terms in the individual bounds by d, and summing over all the documents.

Recall from Section 5.3 that our overall approach to finding empirical Bayes estimates is based on a variational EM procedure. In the variational E-step, discussed in Appendix A.3, we maximize the bound $\mathcal{L}(\gamma, \phi; \alpha, \beta)$ with respect to the variational parameters γ and ϕ. In the M-step, which we describe in this section, we maximize the bound with respect to the model parameters α and β. The overall procedure can thus be viewed as coordinate ascent in $\mathcal{L}$.

A.4.1 CONDITIONAL MULTINOMIALS

To maximize with respect to β, we isolate terms and add Lagrange multipliers:

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} + \sum_{i=1}^{k} \lambda_i \left(\sum_{j=1}^{V} \beta_{ij} - 1 \right).$$

We take the derivative with respect to β_{ij}, set it to zero, and find:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$

A.4.2 DIRICHLET

The terms which contain α are:

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{M} \left(\log \Gamma \left(\sum_{j=1}^{k} \alpha_j \right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} \left((\alpha_i - 1) \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right) \right) \right) \right)$$

Taking the derivative with respect to α_i gives:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = M \left(\Psi \left(\sum_{j=1}^{k} \alpha_j \right) - \Psi(\alpha_i) \right) + \sum_{d=1}^{M} \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right) \right)$$

This derivative depends on α_j, where $j \neq i$, and we therefore must use an iterative method to find the maximal α. In particular, the Hessian is in the form found in Eq. (10):

$$\frac{\partial \mathcal{L}}{\partial \alpha_i \alpha_j} = \delta(i,j) M \Psi'(\alpha_i) - \Psi' \left(\sum_{j=1}^{k} \alpha_j \right),$$

and thus we can invoke the linear-time Newton-Raphson algorithm described in Appendix A.2.

Finally, note that we can use the same algorithm to find an empirical Bayes point estimate of η, the scalar parameter for the exchangeable Dirichlet in the smoothed LDA model in Section 5.4.

Introduction to Probabilistic Topic Models

David M. Blei
Princeton University

Abstract

Probabilistic topic models are a suite of algorithms whose aim is to discover the hidden thematic structure in large archives of documents. In this article, we review the main ideas of this field, survey the current state-of-the-art, and describe some promising future directions. We first describe latent Dirichlet allocation (LDA) [8], which is the simplest kind of topic model. We discuss its connections to probabilistic modeling, and describe two kinds of algorithms for topic discovery. We then survey the growing body of research that extends and applies topic models in interesting ways. These extensions have been developed by relaxing some of the statistical assumptions of LDA, incorporating meta-data into the analysis of the documents, and using similar kinds of models on a diversity of data types such as social networks, images and genetics. Finally, we give our thoughts as to some of the important unexplored directions for topic modeling. These include rigorous methods for checking models built for data exploration, new approaches to visualizing text and other high dimensional data, and moving beyond traditional information engineering applications towards using topic models for more scientific ends.

1 Introduction

As our collective knowledge continues to be digitized and stored—in the form of news, blogs, web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search and understand these vast amounts of information.

Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing.

Imagine searching and exploring documents based on the themes that run through them. We might "zoom in" and "zoom out" to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

For example, consider using themes to explore the complete history of the New York Times. At a broad level some of the themes might correspond to the sections of the newspaper—foreign policy, national affairs, sports. We could zoom in on a theme of interest, such as foreign policy, to reveal various aspects of it—Chinese foreign policy, the conflict in the Middle East, the United States's relationship with Russia. We could then navigate through time to reveal how these specific themes have changed, tracking, for example, the changes in the conflict in the Middle East over the last fifty years. And, in all of this exploration, we would be pointed to the original articles relevant to the themes. The thematic structure would be a new kind of window through which to explore and digest the collection.

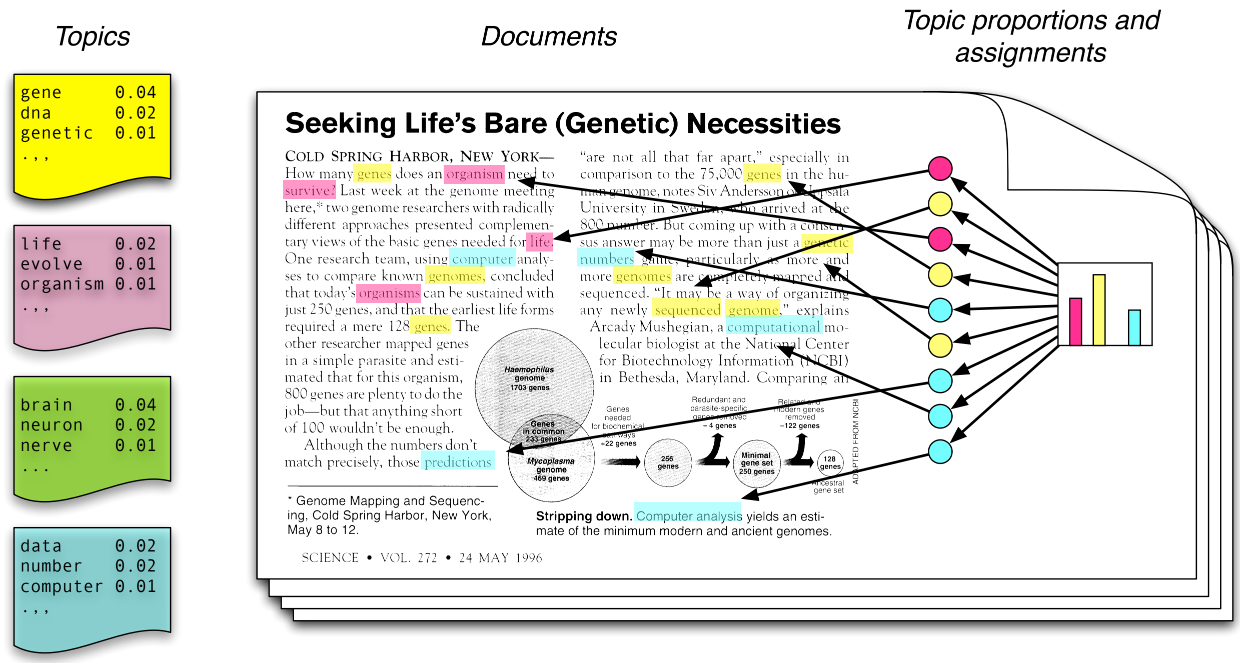
But we don't interact with electronic archives in this way. While more and more texts are available online, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed *probabilistic topic modeling*, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. (See, for example, Figure 3 for topics found by analyzing the *Yale Law Journal*.) Topic modeling algorithms do not require any prior annotations or labeling of the documents—the topics emerge from the analysis of the original texts. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation.

2 Latent Dirichlet allocation

We first describe the basic ideas behind *latent Dirichlet allocation* (LDA), which is the simplest topic model [8]. The intuition behind LDA is that documents exhibit multiple topics. For example, consider the article in Figure 1. This article, entitled "Seeking Life's Bare (Genetic) Necessities," is about using data analysis to determine the number of genes that an organism needs to survive (in an evolutionary sense).

By hand, we have highlighted different words that are used in the article. Words about *data analysis*, such as "computer" and "prediction," are highlighted in blue; words about *evolutionary biology*, such as "life" and "organism", are highlighted in pink; words about *genetics,* such as "sequenced" and "genes," are highlighted in yellow. If we took the time to highlight every word in the article, you would see that this article blends genetics, data analysis, and evolutionary biology with different proportions. (We exclude words, such as "and" "but" or "if," which contain little topical content.) Furthermore, knowing that this article blends those topics would help you situate it in a collection of scientific articles.

LDA is a statistical model of document collections that tries to capture this intuition. It is most easily described by its generative process, the imaginary random process by which the


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
.,,
life 0.02
evolve 0.01
organism 0.01
.,,
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
.,,
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure 1: **The intuitions behind latent Dirichlet allocation.** We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

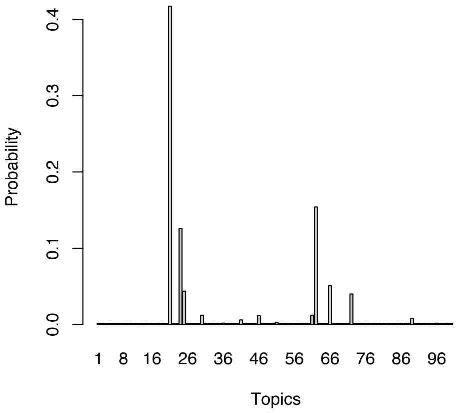
model assumes the documents arose. (The interpretation of LDA as a probabilistic model is fleshed out below in Section 2.1.)

We formally define a *topic* to be a distribution over a fixed vocabulary. For example the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.[1] Now for each document in the collection, we generate the words in a two-stage process.

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics with different proportion (step #1); each word in each document

[1] Technically, the model assumes that the topics are generated first, before the documents.



"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figure 2: **Real inference with LDA.** We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left is the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).[2]

In the example article, the distribution over topics would place probability on *genetics*, *data analysis* and *evolutionary biology*, and each word is drawn from one of those three topics. Notice that the next article in the collection might be about *data analysis* and *neuroscience*; its distribution over topics would place probability on those two topics. This is the distinguishing characteristic of latent Dirichlet allocation—all the documents in the collection share the same set of topics, but each document exhibits those topics with different proportion.

As we described in the introduction, the goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure—the topics, per-document topic distributions, and the per-document per-word topic assignments—are *hidden structure*. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as "reversing" the generative process—what is the hidden structure that likely generated the observed collection?

Figure 2 illustrates example inference using the same example document from Figure 1. Here, we took 17,000 articles from *Science* magazine and used a topic modeling algorithm to infer the hidden topic structure. (The algorithm assumed that there were 100 topics.) We

[2] We should explain the mysterious name, "latent Dirichlet allocation." The distribution that is used to draw the per-document topic distributions in step #1 (the cartoon histogram in Figure 1) is called a *Dirichlet distribution*. In the generative process for LDA, the result of the Dirichlet is used to *allocate* the words of the document to different topics. Why *latent*? Keep reading.

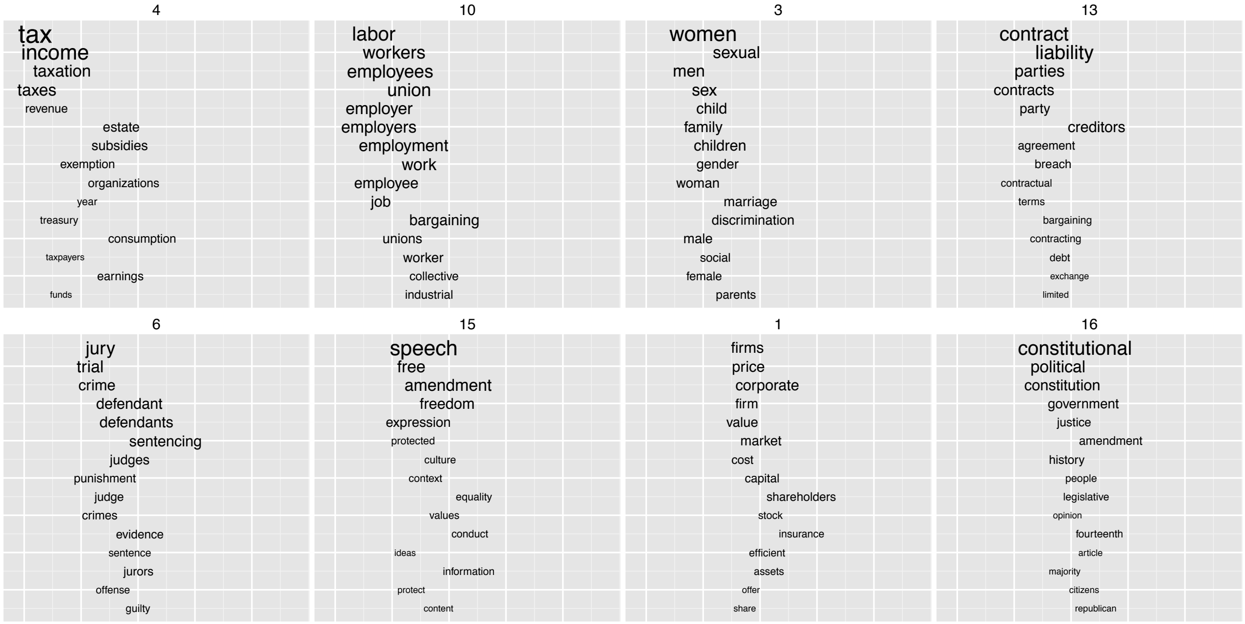


Figure 3: A topic model fit to the *Yale Law Journal*. Here there are twenty topics (the top eight are plotted). Each topic is illustrated with its top most frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax."

then computed the inferred topic distribution for the example article (Figure 2, left), the distribution over topics that best describes its particular collection of words. Notice that this topic distribution, though it can use any of the topics, has only "activated" a handful of them. Further, we can examine the most probable terms from each of the most probable topics (Figure 2, right). On examination, we see that these terms are recognizable as terms about genetics, survival, and data analysis, the topics that are combined in the example article.

We emphasize that the algorithms have no information about these subjects and the articles are not labeled with topics or keywords. The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents.[3] For example, Figure 3 illustrates topics discovered from *Yale Law Journal*. (Here the number of topics was set to be twenty.) Topics about subjects like genetics and data analysis are replaced by topics about discrimination and contract law.

The utility of topic models stems from the property that the inferred hidden structure resembles the thematic structure of the collection. This interpretable hidden structure annotates each document in the collection—a task that is painstaking to perform by hand—and these annotations can be used to aid tasks like information retrieval, classification, and

[3] Indeed calling these models "topic models" is retrospective—the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA.

corpus exploration.[4] In this way, topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts.

2.1 LDA and probabilistic models

LDA and other topic models are part of the larger field of *probabilistic modeling.* In generative probabilistic modeling, we treat our data as arising from a generative process that includes *hidden variables.* This generative process defines a *joint probability distribution* over both the observed and hidden random variables. We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the observed variables. This conditional distribution is also called the *posterior distribution.*

LDA falls precisely into this framework. The observed variables are the words of the documents; the hidden variables are the topic structure; and the generative process is as described above. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of the hidden variables given the documents.

We can describe LDA more formally with the following notation. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary (the distributions over words at left in Figure 1). The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d (the cartoon histogram in Figure 1). The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in document d (the colored coin in Figure 1). Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$\begin{aligned} & p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) \\ & \quad = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\textstyle\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right). \end{aligned} \tag{1}$$

Notice that this distribution specifies a number of dependencies. For example, the topic assignment $z_{d,n}$ depends on the per-document topic proportions θ_d. As another example, the observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ and *all* of the topics $\beta_{1:K}$. (Operationally, that term is defined by looking up which topic $z_{d,n}$ refers to and looking up the probability of the word $w_{d,n}$ within that topic.)

These dependencies define LDA. They are encoded in the statistical assumptions behind the generative process, in the particular mathematical form of the joint distribution, and—in a third way—in the *probabilistic graphical model* for LDA. Probabilistic graphical models provide

[4] See, for example, the browser of *Wikipedia* built with a topic model at `http://www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/browse/topic-list.html`.

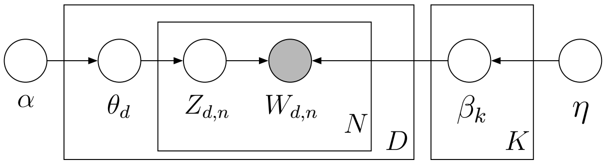


Figure 4: **The graphical model for latent Dirichlet allocation.** Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes–the topic proportions, assignments and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The N plate denotes the collection words within documents; the D plate denotes the collection of documents within the collection.

a graphical language for describing families of probability distributions.[5] The graphical model for LDA is in Figure 4. These three representations are equivalent ways of describing the probabilistic assumptions behind LDA.

In the next section, we describe the inference algorithms for LDA. However, we first pause to describe the short history of these ideas. LDA was developed to fix an issue with a previously developed probabilistic model *probabilistic latent semantic analysis* (pLSI) [21]. That model was itself a probabilistic version of the seminal work on *latent semantic analysis* [14], which revealed the utility of the singular value decomposition of the document-term matrix. From this matrix factorization perspective, LDA can also be seen as a type of principal component analysis for discrete data [11, 12].

2.2 Posterior computation for LDA

We now turn to the computational problem, computing the conditional distribution of the topic structure given the observed documents. (As we mentioned above, this is called the *posterior.*) Using our notation, the posterior is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \,|\, w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \tag{2}$$

The numerator is the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator is the *marginal probability* of the observations, which is the probability of seeing the observed corpus under any topic model. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure.

[5] The field of graphical models is actually more than a language for describing families of distributions. It is a field that illuminates the deep mathematical links between probabilistic independence, graph theory, and algorithms for computing with probability distributions [35].

That number of possible topic structures, however, is exponentially large; this sum is intractable to compute.[6] As for many modern probabilistic models of interest—and for much of modern Bayesian statistics—we cannot compute the posterior because of the denominator, which is known as the *evidence*. A central research goal of modern probabilistic modeling is to develop efficient methods for approximating it. Topic modeling algorithms—like the algorithms used to create Figure 1 and Figure 3—are often adaptations of general-purpose methods for approximating the posterior distribution.

Topic modeling algorithms form an approximation of Equation 2 by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior. Topic modeling algorithms generally fall into two categories—sampling-based algorithms and variational algorithms.

Sampling based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. The most commonly used sampling algorithm for topic modeling is *Gibbs sampling*, where we construct a *Markov chain*—a sequence of random variables, each dependent on the previous—whose limiting distribution is the posterior. The Markov chain is defined on the hidden topic variables for a particular corpus, and the algorithm is to run the chain for a long time, collect samples from the limiting distribution, and then approximate the distribution with the collected samples. (Often, just one sample is collected as an approximation of the topic structure with maximal probability.) See [33] for a good description of Gibbs sampling for LDA, and see `http://CRAN.R-project.org/package=lda` for a fast open-source implementation.

Variational methods are a deterministic alternative to sampling-based algorithms [22, 35]. Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.[7] Thus, the inference problem is transformed to an optimization problem. Variational methods open the door for innovations in optimization to have practical impact in probabilistic modeling. See [8] for a coordinate ascent variational inference algorithm for LDA; see [20] for a much faster online algorithm (and open-source software) that easily handles millions of documents and can accommodate streaming collections of text.

Loosely speaking, both types of algorithms perform a search over the topic structure. The collection of documents (the observed random variables in the model) are held fixed and serve as a guide towards where to search. Which approach is better depends on the particular topic model being used—we have so far focused on LDA, but see below for other topic models—and is a source of academic debate. For a good discussion of the merits and drawbacks of both, see [1].

[6] More technically, the sum is over all possible ways of assigning each observed word of the collection to one of the topics. Document collections usually contain observed words at least on the order of millions.

[7] Closeness is measured with *Kullback-Leibler divergence*, an information theoretic measurement of the distance between two probability distributions.

3 Research in topic modeling

The simple LDA model provides a powerful tool for discovering and exploiting the hidden thematic structure in large archives of text. However, one of the main advantages of formulating LDA as a probabilistic model is that it can easily be used as a module in more complicated models for more complicated goals. Since its introduction, LDA has been extended and adapted in many ways.

3.1 Relaxing the assumptions of LDA

LDA is defined by the statistical assumptions it makes about the corpus. One active area of topic modeling research is how to relax and extend these assumptions to uncover more sophisticated structure in the texts.

One assumption that LDA makes is the "bag of words" assumption, that the order of the words in the document does not matter. (To see this note that the joint distribution of Equation 1 remains invariant to permutation of the words of the documents.) While this assumption is unrealistic, it is reasonable if our only goal is to uncover the course semantic structure of the texts.[8] For more sophisticated goals—such as language generation—it is patently not appropriate. There have been a number of extensions to LDA that model words nonexchangeably. For example, [36] developed a topic model that relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word; [18] developed a topic model that switches between LDA and a standard HMM. These models expand the parameter space significantly, but show improved language modeling performance.

Another assumption is that the order of documents does not matter. Again, this can be seen by noticing that Equation 1 remains invariant to permutations of the ordering of documents in the collection. This assumption may be unrealistic when analyzing long-running collections that span years or centuries. In such collections we may want to assume that the *topics* change over time. One approach to this problem is the dynamic topic model [5]—a model that respects the ordering of the documents and gives a richer posterior topical structure than LDA. Figure 5 shows a topic that results from analyzing all of *Science* magazine under the dynamic topic model. Rather than a single distribution over words, a topic is now a sequence of distributions over words. We can find an underlying theme of the collection and track how it has changed over time.

A third assumption about LDA is that the number of topics is assumed known and fixed. The Bayesian nonparametric topic model [34] provides an elegant solution: The number of topics is determined by the collection during posterior inference, and furthermore new documents can exhibit previously unseen topics. Bayesian nonparametric topic models have been extended to hierarchies of topics, which find a tree of topics, moving from more general to more concrete, whose particular structure is inferred from the data [3].

[8] As a thought experiment, imagine shuffling the words of the article in Figure 1. Even when shuffled, you would be able to glean that the article has something to do with genetics.

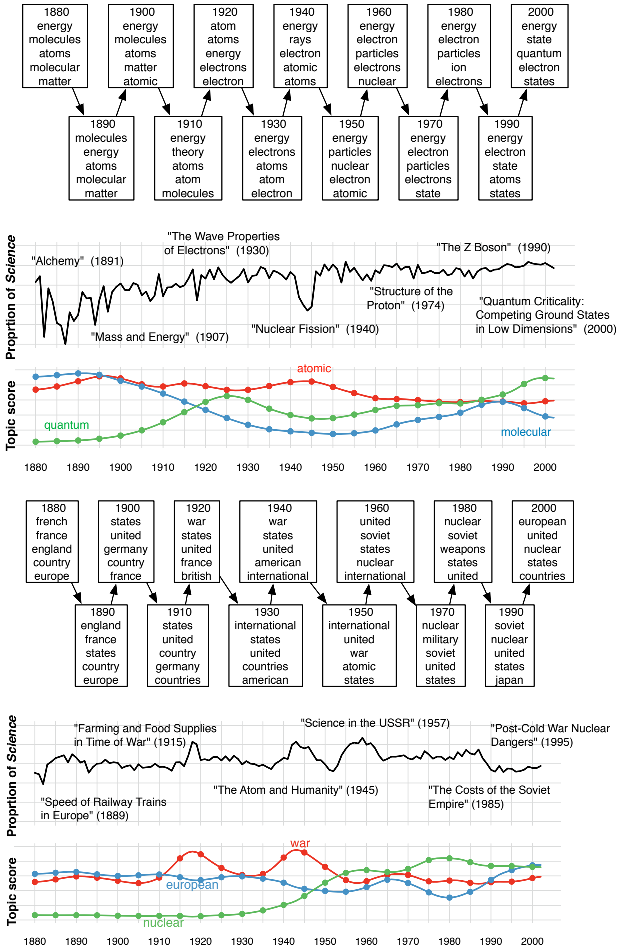


Figure 5: **Two topics from a dynamic topic model.** This model was fit to *Science* from (1880–2002). We have illustrated the top words at each decade.

There are still other extensions of LDA that relax various assumptions made by the model. The correlated topic model [6] and pachinko allocation machine [24] allow the occurrence of topics to exhibit correlation (for example a document about *geology* is more likely to also be about *chemistry* then it is to be about *sports*); the spherical topic model [28] allows words to be *unlikely* in a topic (for example, “wrench” will be particularly unlikely in a topic about *cats*); sparse topic models enforce further structure in the topic distributions [37]; and “bursty” topic models provide a more realistic model of word counts [15].

3.2 Incorporating meta-data

In many text analysis settings, the documents contain additional information—such as author, title, geographic location, links, and others—that we might want to account for when fitting a topic model. There has been a flurry of research on adapting topic models to include meta-data.

The author-topic model [29] is an early success story for this kind of research. The topic proportions are attached to authors; papers with multiple authors are assumed to attach each word to an author, drawn from a topic drawn from his or her topic proportions. The author-topic model allows for inferences about authors as well as documents. Rosen-Zvi et al. show examples of author similarity based on their topic proportions—such computations are not possible with LDA.

Many document collections are linked—for example scientific papers are linked by citation or web pages are linked by hyperlink—and several topic models have been developed to account for those links when estimating the topics. The *relational topic model* of [13] assumes that each document is modeled as in LDA and that the links between documents depend on the distance between their topic proportions. This is both a new topic model and a new network model. Unlike traditional statistical models of networks, the relational topic model takes into account node attributes (here, the words of the documents) in modeling the links.

Other work that incorporates meta-data into topic models includes models of linguistic structure [10], models that account for distances between corpora [38], and models of named entities [26]. General purpose methods for incorporating meta-data into topic models include Dirichlet-multinomial regression models [25] and supervised topic models [7].

3.3 Other kinds of data

In LDA, the topics are distributions over words and this discrete distribution generates observations (words in documents). One advantage of LDA is that these choices for the topic parameter and data-generating distribution can be adapted to other kinds of observations with only small changes to the corresponding inference algorithms. As a class of models, LDA can be thought of as a *mixed-membership model* of grouped data—rather than associate each group of observations (document) with one component (topic), each group exhibits multiple components with different proportions. LDA-like models have been adapted to many kinds of data, including survey data, user preferences, audio and music, computer code, network logs, and social networks. We describe two areas where mixed-membership models have been particularly successful.

In population genetics, the same probabilistic model was independently invented to find ancestral populations (e.g., originating from Africa, Europe, the Middle East, etc.) in the genetic ancestry of a sample of individuals [27]. The idea is that each individual's genotype descends from one or more of the ancestral populations. Using a model much like LDA, biologists can both characterize the genetic patterns in those populations (the "topics") and identify how each individual expresses them (the "topic proportions"). This model is powerful because the genetic patterns in ancestral populations can be hypothesized, even when "pure" samples from them are not available.

LDA has been widely used and adapted in computer vision, where the inference algorithms are applied to natural images in the service of image retrieval, classification, and organization. Computer vision researchers have made a direct analogy from images to documents. In document analysis we assume that documents exhibit multiple topics and a collection of documents exhibits the same set of topics. In image analysis we assume that each image exhibits a combination of visual patterns and that the same visual patterns recur throughout a collection of images. (In a preprocessing step, the images are analyzed to form collections of "visual words.") Topic modeling for computer vision has been used to classify images [16], connect images and captions [4], build image hierarchies [2, 23, 31] and other applications.

4 Future directions

Topic modeling is an emerging field in machine learning, and there are many exciting new directions for research.

Evaluation and model checking. There is a disconnect between how topic models are evaluated and why we expect topic models are useful. Typically, topic models are evaluated in the following way. First, hold out a subset of your corpus as the test set. Then, fit a variety of topic models to the rest of the corpus and approximate a measure of model fit (e.g., probability) for each trained model on the test set. Finally, choose the the model that achieves the best held out performance.

But topic models are often used to organize, summarize and help users explore large corpora, and there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation. One open direction for topic modeling is to develop evaluation methods that match how the algorithms are used. How can we compare topic models based on how interpretable they are?

This is the *model checking* problem. When confronted with a new corpus and a new task, which topic model should I use? How can I decide which of the many modeling assumptions are important for my goals? How should I move between the many kinds of topic models that have been developed? These questions have been given some attention by statisticians [9, 30], but they have been scrutinized less for the scale of problems that machine learning tackles. New computational answers to these questions would be a significant contribution to topic modeling.

Visualization and user interfaces. Another promising future direction for topic modeling is to develop new methods of interacting with and visualizing topics and corpora. Topic models provide new exploratory structure in large collections—how can we best exploit that structure to aid in discovery and exploration?

One problem is how to display the topics. Typically, we display topics by listing the most frequent words of each (see Figure 2), but new ways of labeling the topics—either by choosing different words or displaying the chosen words differently—may be more effective. A further problem is how to best display a document with a topic model. At the document-level, topic models provide potentially useful information about the structure of the document. Combined with effective topic labels, this structure could help readers identify the most interesting parts of the document. Moreover, the hidden topic proportions implicitly connect each document to the other documents (by considering a distance measure between topic proportions). How can we best display these connections? What is an effective interface to the whole corpus and its inferred topic structure?

These are user interface questions, and they are essential to topic modeling. Topic modeling algorithms show much promise for uncovering meaningful thematic structure in large collections of documents. But making this structure useful requires careful attention to information visualization and the corresponding user interfaces.

Topic models for data discovery. Topic models have been developed with information engineering applications in mind. As a statistical model, however, topic models should be able to tell us something, or help us form a hypothesis, about the data. What can we *learn* about the language (and other data) based on the topic model posterior? Some work in this area has appeared in political science [19], bibliometrics [17] and psychology [32]. This kind of research adapts topic models to measure an external variable of interest, a difficult task for unsupervised learning which must be carefully validated.

In general, this problem is best addressed by teaming computer scientists with other scholars to use topic models to help explore, visualize and draw hypotheses from their data. In addition to scientific applications, such as genetics or neuroscience, one can imagine topic models coming to the service of history, sociology, linguistics, political science, legal studies, comparative literature, and other fields where texts are a primary object of study. By working with scholars in diverse fields, we can begin to develop a new interdisciplinary computational methodology for working with and drawing conclusions from archives of texts.

5 Summary

We have surveyed *probabilistic topic models*, a suite of algorithms that provide a statistical solution to the problem of managing large archives of documents. With recent scientific advances in support of unsupervised machine learning—flexible components for modeling, scalable algorithms for posterior inference, and increased access to massive data sets—topic models promise to be an important component for summarizing and understanding our growing digitized archive of information.

References

[1] A. Asuncion, M. Welling, P. Smyth, and Y. Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, 2009.

[2] E. Bart, M. Welling, and P. Perona. Unsupervised organization of image collections: Unsupervised organization of image collections: Taxonomies and beyond. *Transactions on Pattern Recognition and Machine Intelligence*, 2010.

[3] D. Blei, T. Griffiths, and M. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134. ACM Press, 2003.

[5] D. Blei and J. Lafferty. Dynamic topic models. In *International Conference on Machine Learning*, pages 113–120, New York, NY, USA, 2006. ACM.

[6] D. Blei and J. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[7] D. Blei and J. McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, 2007.

[8] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[9] G. Box. Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):383–430, 1980.

[10] J. Boyd-Graber and D. Blei. Syntactic topic models. In *Neural Information Processing Systems*, 2009.

[11] W. Buntine. Variational extentions to EM and multinomial PCA. In *European Conference on Machine Learning*, 2002.

[12] W. Buntine and A. Jakulin. Discrete component analysis. In *Subspace, Latent Structure and Feature Selection*. Springer, 2006.

[13] J. Chang and D. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1), 2010.

[14] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[15] G. Doyle and C. Elkan. Accounting for burstiness in topic models. In *International Conference on Machine Learning*, pages 281–288. ACM, 2009.

[16] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *IEEE Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[17] S. Gerrish and D. Blei. A language-based approach to measuring scholarly impact. In *International Conference on Machine Learning*, 2010.

[18] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537–544, Cambridge, MA, 2005. MIT Press.

[19] J. Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1, 2010.

[20] M. Hoffman, D. Blei, and F. Bach. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems*, 2010.

[21] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence (UAI)*, 1999.

[22] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[23] J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual image hierarchy. In *Computer Vision and Pattern Recognition*, 2010.

[24] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *International Conference on Machine Learning*, pages 577–584, 2006.

[25] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, 2008.

[26] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *Knowledge Discovery and Data Mining*, 2006.

[27] J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, June 2000.

[28] J. Reisinger, A. Waters, B. Silverthorn, and R. Mooney. Spherical topic models. In *International Conference on Machine Learning*, 2010.

[29] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smith. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.

[30] D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

[31] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. Unsupervised discovery of visual object class hierarchies. In *Conference on Computer Vision and Pattern Recognition*, 2008.

[32] R. Socher, S. Gershman, A. Perotte, P. Sederberg, D. Blei, and K. Norman. A Bayesian analysis of dynamics in free recall. In *Neural Information Processing Systems*, 2009.

[33] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006.

[34] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[35] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[36] H. Wallach. Topic modeling: Beyond bag of words. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[37] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1982–1989. 2009.

[38] C. Wang, B. Thiesson, C. Meek, and D. Blei. Markov topic models. In *Artificial Intelligence and Statistics*, 2009.

Probabilistic Topic Models

David M. Blei

Department of Computer Science
Princeton University

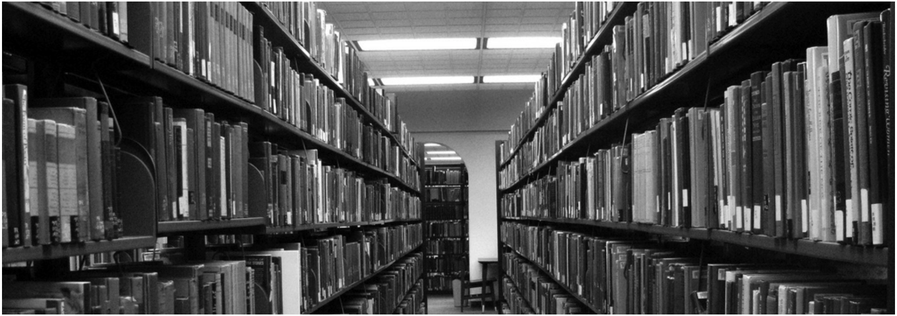
September 2, 2012

Probabilistic topic models

As more information becomes available, it becomes more difficult to find and discover what we need.

We need new tools to help us organize, search, and understand these vast amounts of information.

Probabilistic topic models



Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, search, form predictions.

Probabilistic topic models

topics in "Corpus"
Collection of Docs

topic = distribution of words
collection

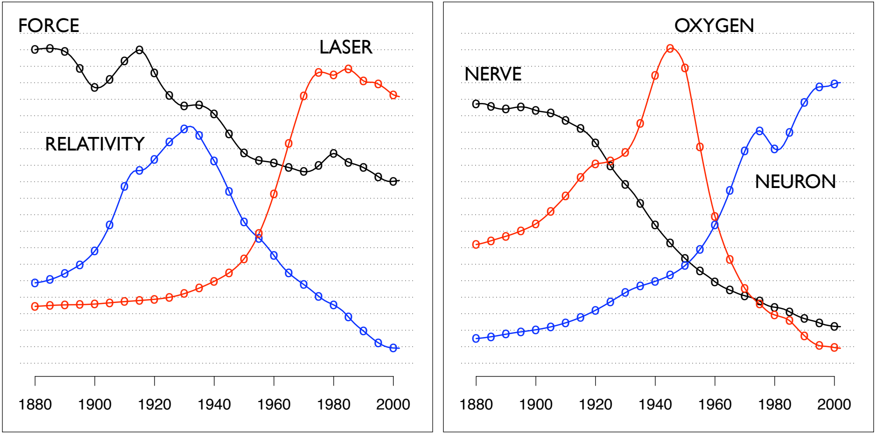
human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

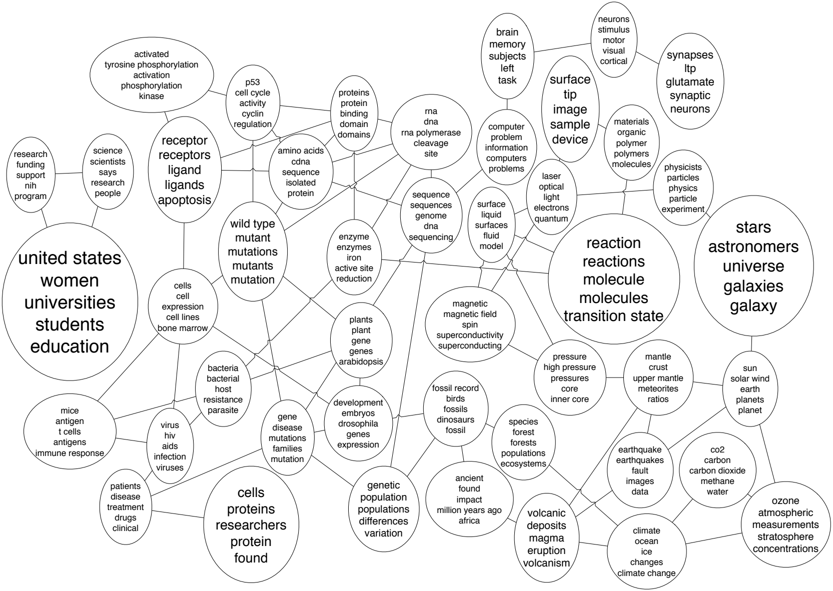
Probabilistic topic models



"Theoretical Physics"

"Neuroscience"

Probabilistic topic models



brain
memory
subjects
left
task

neurons
stimulus
motor
visual
cortical

synapses
ltp
glutamate
synaptic
neurons

activated
tyrosine phosphorylation
activation
phosphorylation
kinase

p53
cell cycle
activity
cyclin
regulation

proteins
protein
binding
domain
domains

rna
dna
rna polymerase
cleavage
site

surface
tip
image
sample
device

materials
organic
polymer
polymers
molecules

computer
problem
information
computers
problems

research
funding
support
nih
program

science
scientists
says
research
people

receptor
receptors
ligand
ligands
apoptosis

amino acids
cdna
sequence
isolated
protein

laser
optical
light
electrons
quantum

physicists
particles
physics
particle
experiment

sequence
sequences
genome
dna
sequencing

surface
liquid
surfaces
fluid
model

wild type
mutant
mutations
mutants
mutation

enzyme
enzymes
iron
active site
reduction

reaction
reactions
molecule
molecules
transition state

stars
astronomers
universe
galaxies
galaxy

united states
women
universities
students
education

cells
cell
expression
cell lines
bone marrow

magnetic
magnetic field
spin
superconductivity
superconducting

plants
plant
gene
genes
arabidopsis

pressure
high pressure
pressures
core
inner core

mantle
crust
upper mantle
meteorites
ratios

sun
solar wind
earth
planets
planet

bacteria
bacterial
host
resistance
parasite

fossil record
birds
fossils
dinosaurs
fossil

development
embryos
drosophila
genes
expression

mice
antigen
t cells
antigens
immune response

virus
hiv
aids
infection
viruses

gene
disease
mutations
families
mutation

species
forest
forests
populations
ecosystems

earthquake
earthquakes
fault
images
data

co2
carbon
carbon dioxide
methane
water

patients
disease
treatment
drugs
clinical

cells
proteins
researchers
protein
found

genetic
population
populations
differences
variation

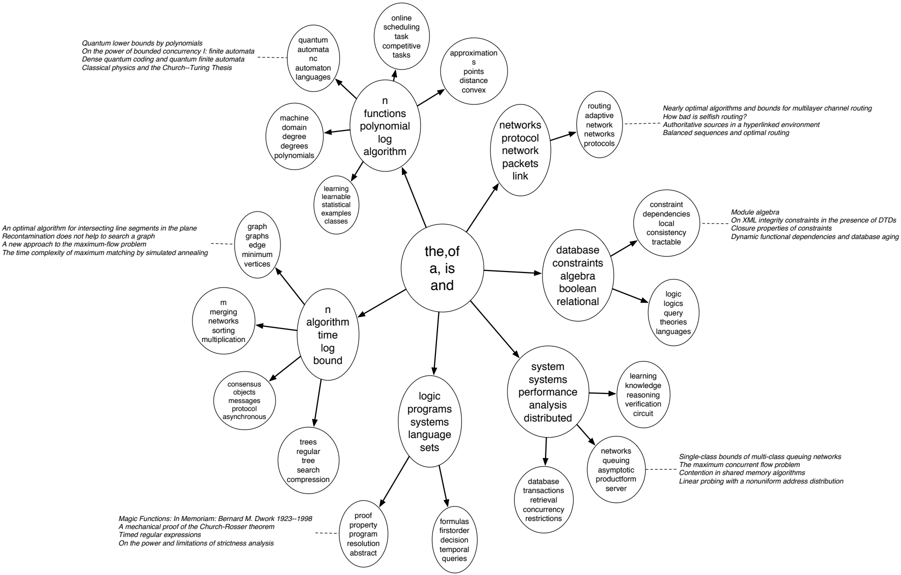
ancient
found
impact
million years ago
africa

volcanic
deposits
magma
eruption
volcanism

climate
ocean
ice
changes
climate change

ozone
atmospheric
measurements
stratosphere
concentrations

Probabilistic topic models


the,of a, is and
n functions polynomial log algorithm
quantum automata nc automaton languages
online scheduling task competitive tasks
approximations points distance convex
machine domain degree degrees polynomials
learning learnable statistical examples classes
networks protocol network packets link
routing adaptive network networks protocols
database constraints algebra boolean relational
constraint dependencies local consistency tractable
logic logics query theories languages
system systems performance analysis distributed
learning knowledge reasoning verification circuit
networks queuing asymptotic productform server
database transactions retrieval concurrency restrictions
logic programs systems language sets
proof property program resolution abstract
formulas firstorder decision temporal queries
n algorithm time log bound
graph graphs edge minimum vertices
m merging networks sorting multiplication
consensus objects messages protocol asynchronous
trees regular tree search compression

Quantum lower bounds by polynomials
On the power of bounded concurrency I: finite automata
Dense quantum coding and quantum finite automata
Classical physics and the Church--Turing Thesis

Nearly optimal algorithms and bounds for multilayer channel routing
How bad is selfish routing?
Authoritative sources in a hyperlinked environment
Balanced sequences and optimal routing

Module algebra
On XML integrity constraints in the presence of DTDs
Closure properties of constraints
Dynamic functional dependencies and database aging

An optimal algorithm for intersecting line segments in the plane
Recontamination does not help to search a graph
A new approach to the maximum-flow problem
The time complexity of maximum matching by simulated annealing

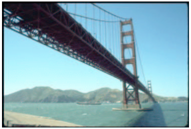
Single-class bounds of multi-class queuing networks
The maximum concurrent flow problem
Contention in shared memory algorithms
Linear probing with a nonuniform address distribution

Magic Functions: In Memoriam: Bernard M. Dwork 1923--1998
A mechanical proof of the Church-Rosser theorem
Timed regular expressions
On the power and limitations of strictness analysis

Probabilistic topic models

SKY WATER TREE MOUNTAIN PEOPLE

SCOTLAND WATER FLOWER HILLS TREE



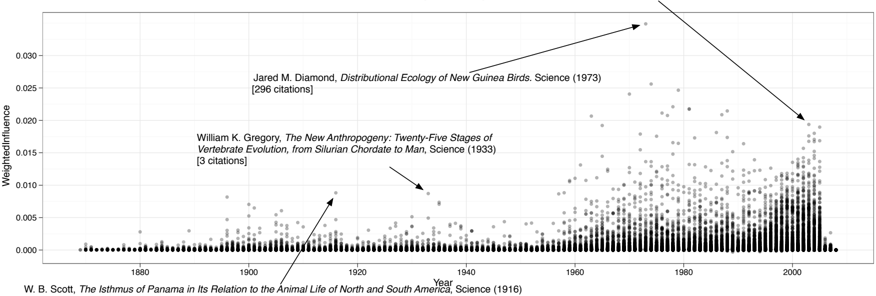
SKY WATER BUILDING PEOPLE WATER

FISH WATER OCEAN TREE CORAL

PEOPLE MARKET PATTERN TEXTILE DISPLAY

BIRDS NEST TREE BRANCH LEAVES

Probabilistic topic models



Derek E. Wildman et al., Implications of Natural Selection in Shaping 99.4% Nonsynonymous DNA Identity between Humans and Chimpanzees: Enlarging Genus Homo, PNAS (2003)
[178 citations]

Jared M. Diamond, *Distributional Ecology of New Guinea Birds.* Science (1973)
[296 citations]

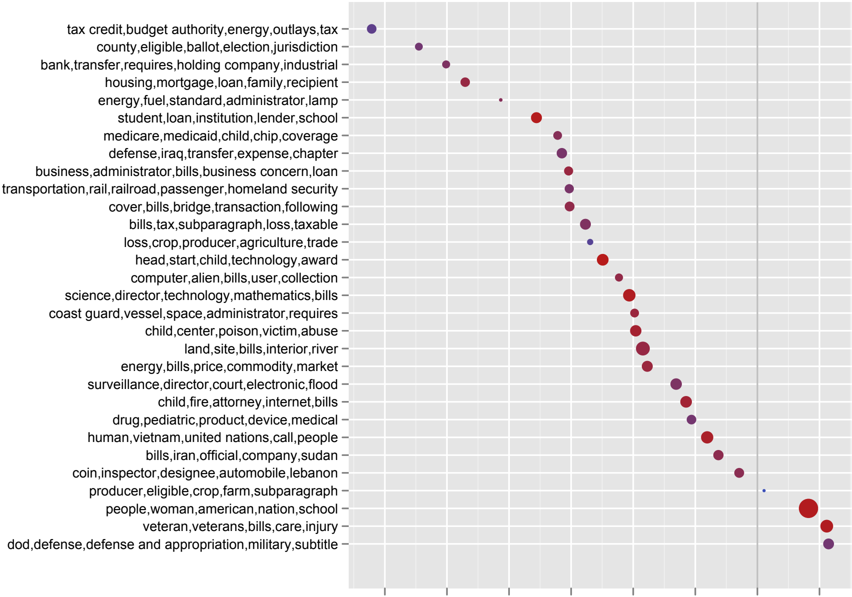
William K. Gregory, *The New Anthropogeny: Twenty-Five Stages of Vertebrate Evolution, from Silurian Chordate to Man*, Science (1933)
[3 citations]

W. B. Scott, *The Isthmus of Panama in Its Relation to the Animal Life of North and South America*, Science (1916)
[3 citations]

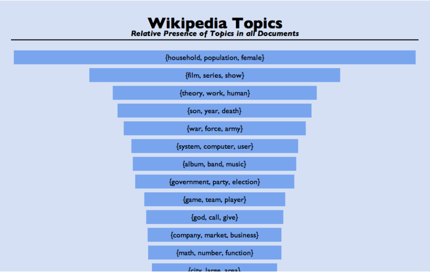
Probabilistic topic models

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo	RTM (ψ_e)
Rates of convergence of the Hastings and Metropolis algorithms	
Possible biases induced by MCMC convergence diagnostics	
Bounding convergence time of the Gibbs sampler in Bayesian image restoration	
Self regenerative Markov chain Monte Carlo	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation	
Diagnosing convergence of Markov chain Monte Carlo algorithms	
Exact Bound for the Convergence of Metropolis Chains	**LDA + Regression**
Self regenerative Markov chain Monte Carlo	
Minorization conditions and convergence rates for Markov chain Monte Carlo	
Gibbs-markov models	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models	
Mediating instrumental variables	
A qualitative framework for probabilistic inference	
Adaptation for Self Regenerative MCMC	

Probabilistic topic models



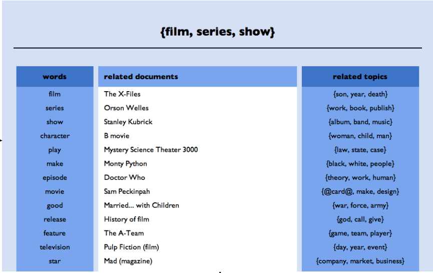
Probabilistic topic models



Wikipedia Topics

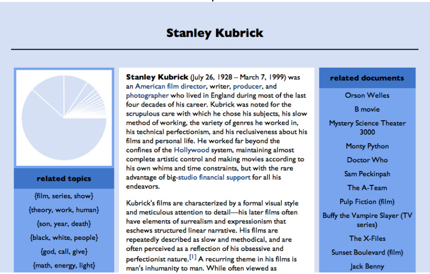
Relative Presence of Topics in all Documents

- {household, population, female}
- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {war, force, army}
- {system, computer, user}
- {album, band, music}
- {government, party, election}
- {game, team, player}
- {god, call, give}
- {company, market, business}
- {math, number, function}
- {city, large, area}



{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}



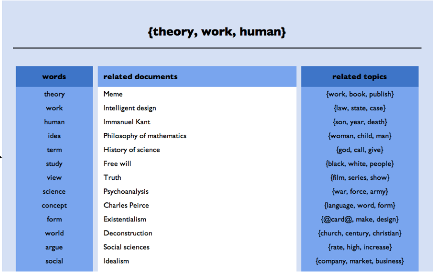
Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related topics
{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
{math, energy, light}

related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)
Jack Benny



{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

Probabilistic topic models

- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I evaluate and check a topic model?
- What are some unanswered questions in this field?
- How can I learn more?

Probabilistic models

- This is a case study in **data analysis with probability models**.
- Our agenda is to teach about this kind of analysis *through* topic models.
- Note: We are being “Bayesian” in this sense:

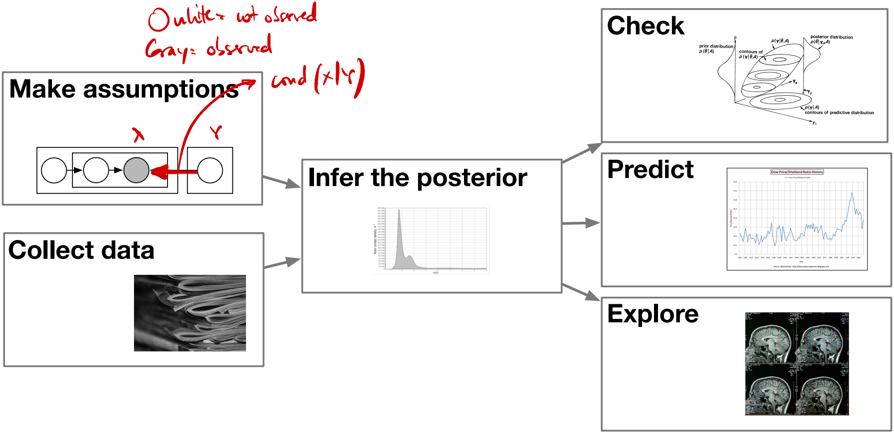
  “[By Bayesian inference,] I simply mean the method of statistical inference that draws conclusions by calculating conditional distributions of unknown quantities given (a) known quantities and (b) model specifications.” (Rubin, 1984)

- (The Bayesian versus Frequentist debate is not relevant to this talk.)

Probabilistic models

- **Specifying models**
  - Directed graphical models
  - Conjugate priors and nonconjugate priors
  - Time series modeling
  - Hierarchical methods
  - Mixed-membership models
  - Prediction from sparse and noisy inputs
- **Model selection and Bayesian nonparametric methods**
- **Approximate posterior inference**
  - MCMC → Gibbs Sampling
  - Variational inference → optm^n / EM
- **Using and evaluating models**
  - Exploring, describing, summarizing, visualizing data
  - Evaluating model fitness

Probabilistic models



Make assumptions

Collect data

Infer the posterior

Check

Predict

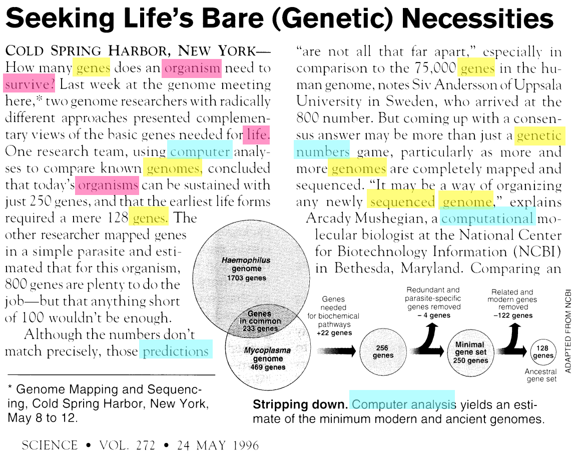
Explore

Organization of these lectures

1. **Introduction to topic modeling: Latent Dirichlet allocation**
2. **Beyond latent Dirichlet allocation**
    - Correlated and dynamic models
    - Supervised models
    - Modeling text and user data
3. **Bayesian nonparametrics: A brief tutorial**
4. **Posterior computation**
    - Scalable variational inference
    - Nonconjugate variational inference
5. **Checking and evaluating models**
    - Using the predictive distribution
    - Posterior predictive checks
6. **Discussion, open questions, and resources**

Introduction to Topic Modeling

Latent Dirichlet allocation (LDA)



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

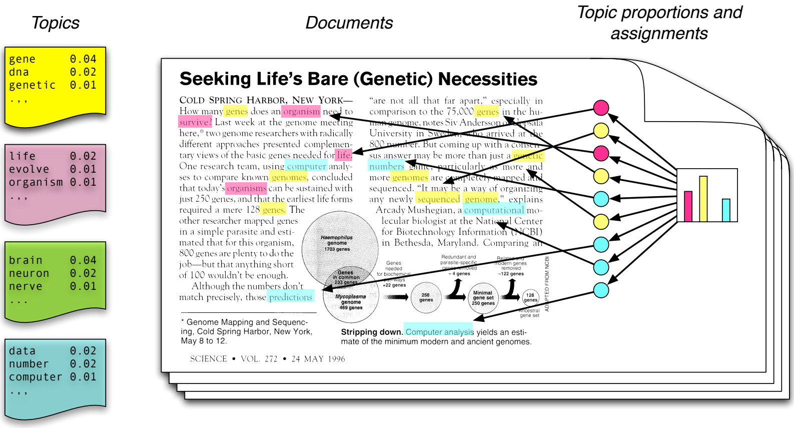
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

Latent Dirichlet allocation (LDA)



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

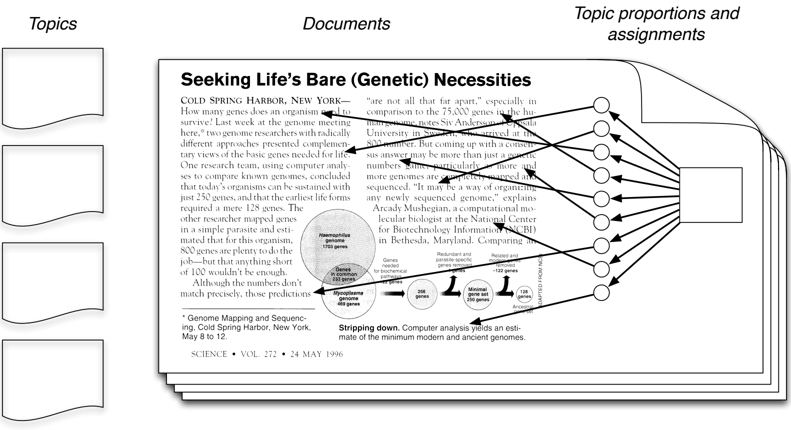
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Latent Dirichlet allocation (LDA)



Topics

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed -4 genes
Minimal gene set 250 genes
Related and modern genes removed -122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

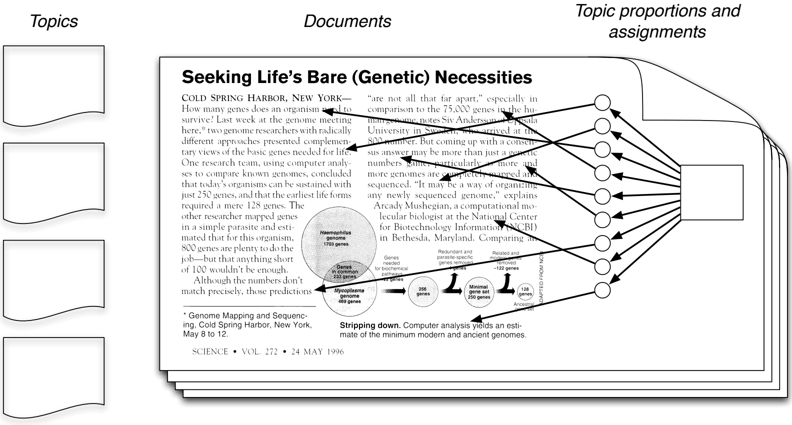
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- In reality, we only observe the documents / words with counts
- The other structure are **hidden variables**

Latent Dirichlet allocation (LDA)



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed -4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

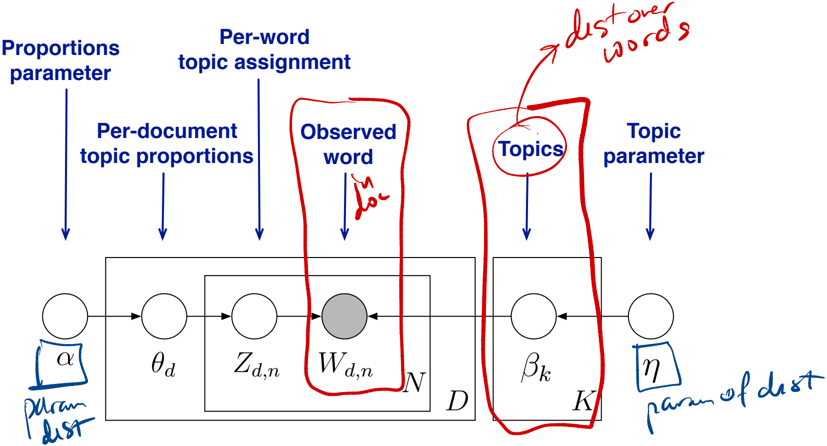
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

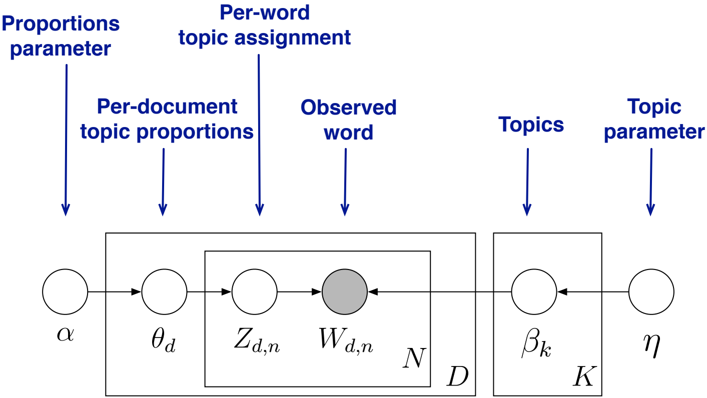
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



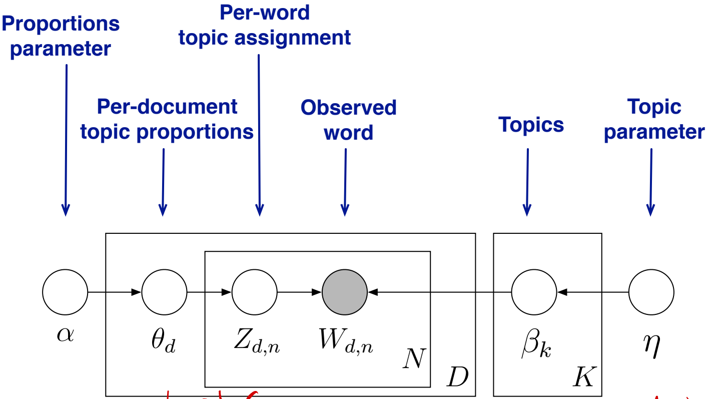
- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

LDA as a graphical model



- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

LDA as a graphical model



Proportions parameter

Per-document topic proportions

Per-word topic assignment

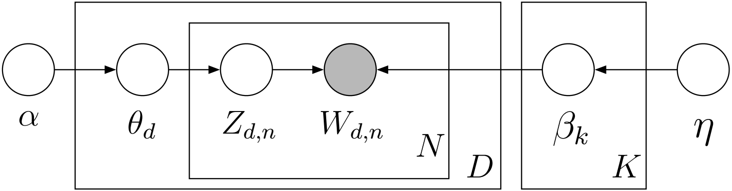
Observed word

Topics

Topic parameter

$$\underbrace{p(\beta,\theta,\mathbf{z},\mathbf{w})}_{\text{posterior likelihood (data fits model)}} = \left(\overbrace{\prod_{i=1}^{K} p(\beta_i\,|\,\eta)}^{\text{topics}}\right)\left(\underbrace{\prod_{d=1}^{D} p(\theta_d\,|\,\alpha)}_{\text{doc prob/topic}}\underbrace{\prod_{n=1}^{N} p(z_{d,n}\,|\,\theta_d)}_{\text{words in topic}}\overbrace{p(w_{d,n}\,|\,\beta_{1:K},z_{d,n})}^{\text{words in doc | topic}}\right)$$

LDA as a graphical model

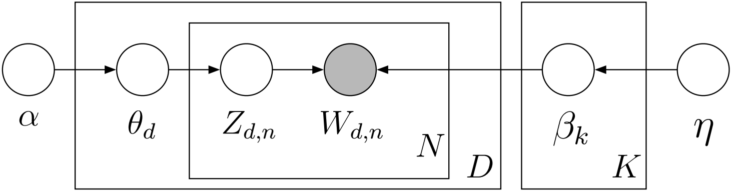


- This joint defines a posterior, $p(\theta, z, \beta \mid w)$.

- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d → prob(doc | topic)
  - Per-corpus topic distributions β_k → prior (topic)

- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.

LDA as a graphical model



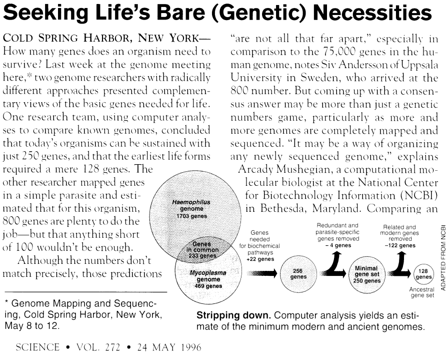
Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)
- Factorization based inference (Arora et al., 2012; Anandkumar et al., 2012)

Example inference

- **Data**: The OCR'ed collection of *Science* from 1990–2000
    - 17K documents
    - 11M words
    - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

Example inference



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

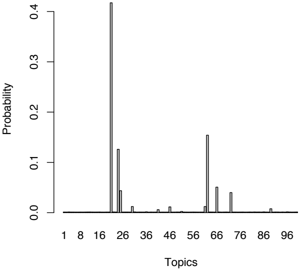
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

→



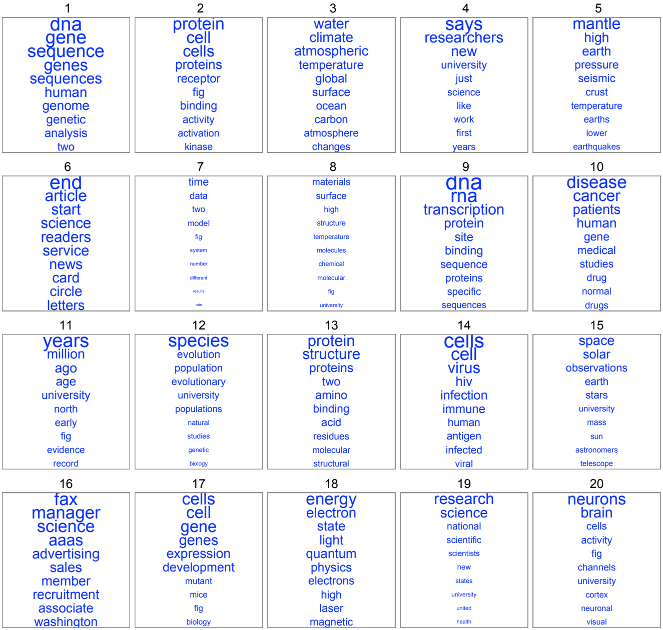
Example inference

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations


1
dna gene sequence genes sequences human genome genetic analysis two
2
protein cell cells proteins receptor fig binding activity activation kinase
3
water climate atmospheric temperature global surface ocean carbon atmosphere changes
4
says researchers new university just science like work first years
5
mantle high earth pressure seismic crust temperature earths lower earthquakes
6
end article start science readers service news card circle letters
7
time data two model fig system number different results rate
8
materials surface high structure temperature molecules chemical molecular fig university
9
dna rna transcription protein site binding sequence proteins specific sequences
10
disease cancer patients human gene medical studies drug normal drugs
11
years million ago age university north early fig evidence record
12
species evolution population evolutionary university populations natural studies genetic biology
13
protein structure proteins two amino binding acid residues molecular structural
14
cells cell virus hiv infection immune human antigen infected viral
15
space solar observations earth stars university mass sun astronomers telescope
16
fax manager science aaas advertising sales member recruitment associate washington
17
cells cell gene genes expression development mutant mice fig biology
18
energy electron state light quantum physics electrons high laser magnetic
19
research science national scientific scientists new states university united health
20
neurons brain cells activity fig channels university cortex neuronal visual

Aside: The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

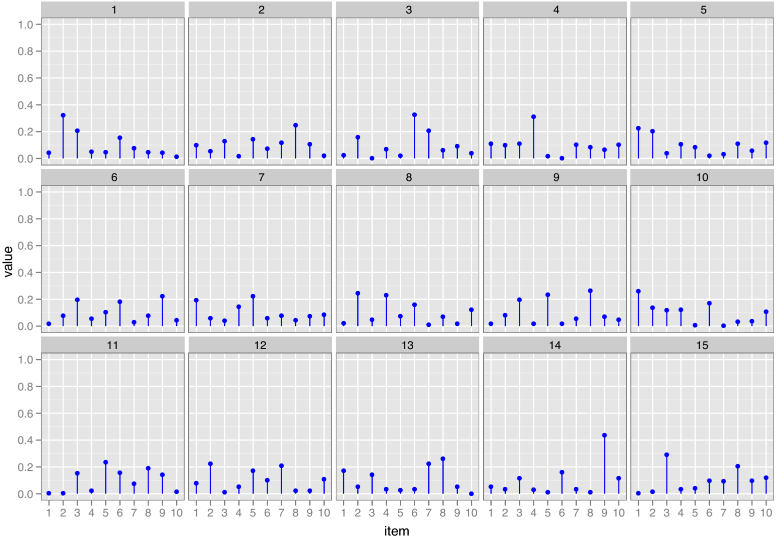
$$p(\theta \,|\, \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.

- The parameter α controls the mean shape and sparsity of θ.

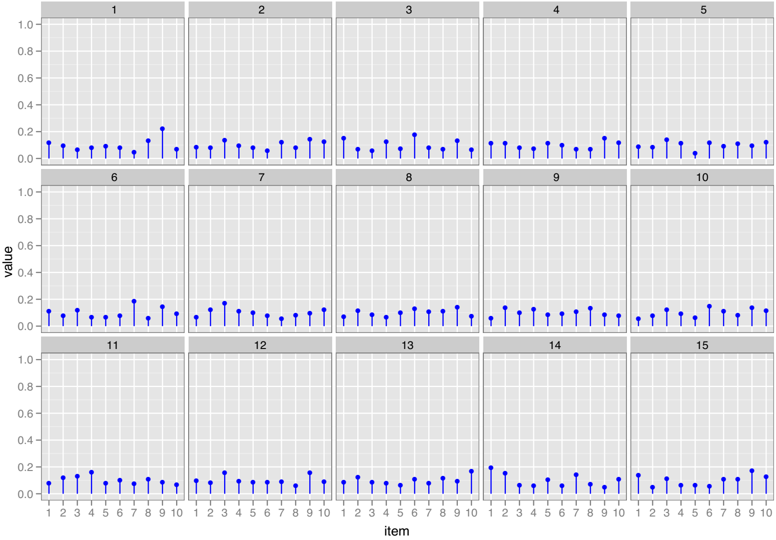
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

$\alpha = 1$

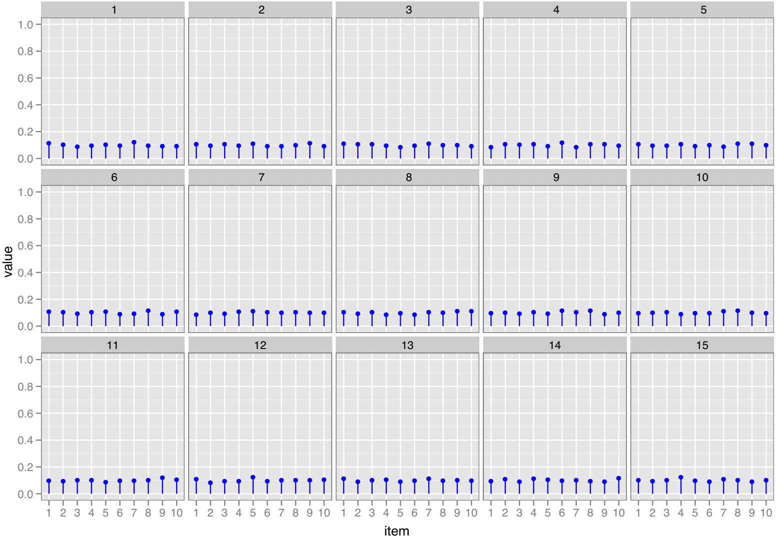


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

$\alpha = 10$

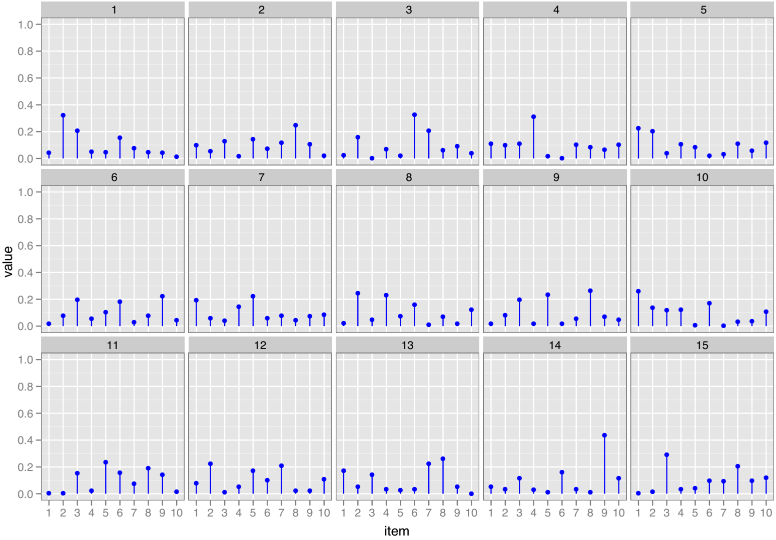

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
value
1 2 3 4 5 6 7 8 9 10
item

$\alpha = 100$

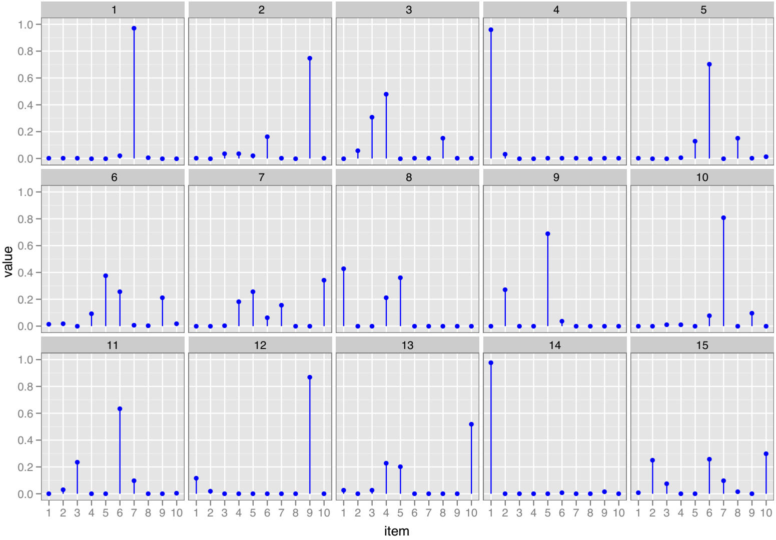


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1
2
3
4
5
6
7
8
9
10

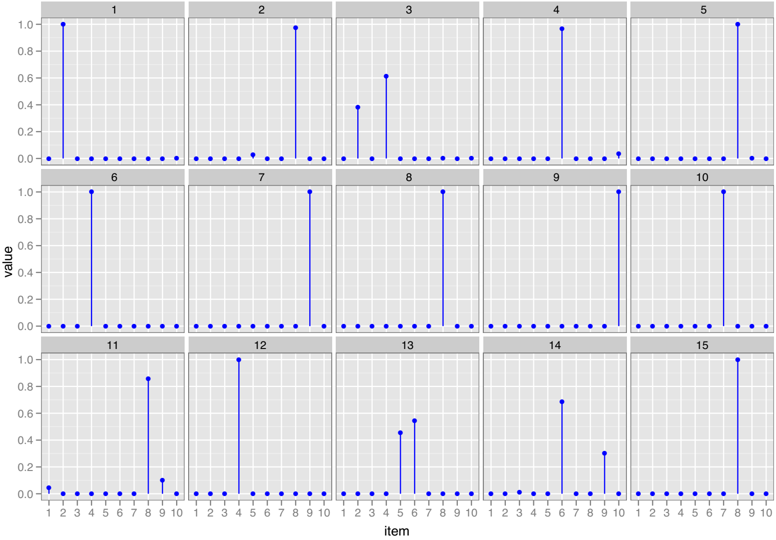
$\alpha = 1$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
value
item
1 2 3 4 5 6 7 8 9 10

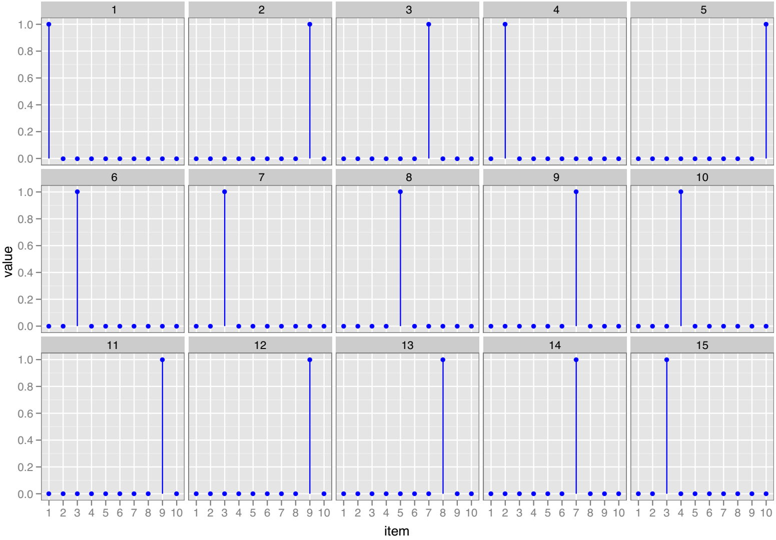
$\alpha = 0.1$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
value
1 2 3 4 5 6 7 8 9 10
item

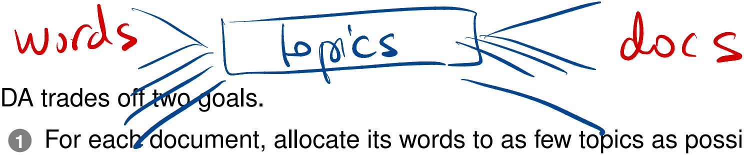
$\alpha = 0.01$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
0.0
0.2
0.4
0.6
0.8
1.0
item
1
2
3
4
5
6
7
8
9
10

$\alpha = 0.001$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1
2
3
4
5
6
7
8
9
10

Why does LDA "work"?



- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few terms as possible.

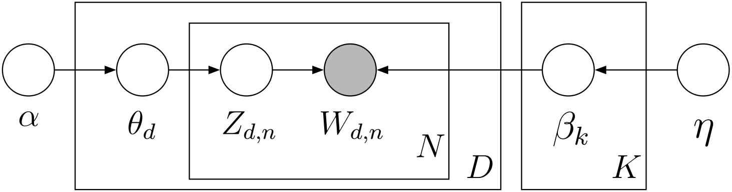
- These goals are at odds.

  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.

  - Putting very few words in each topic makes #1 hard:
    To cover a document's words, it must assign many topics to it.

- Trading off these goals finds groups of tightly co-occurring words.

LDA summary

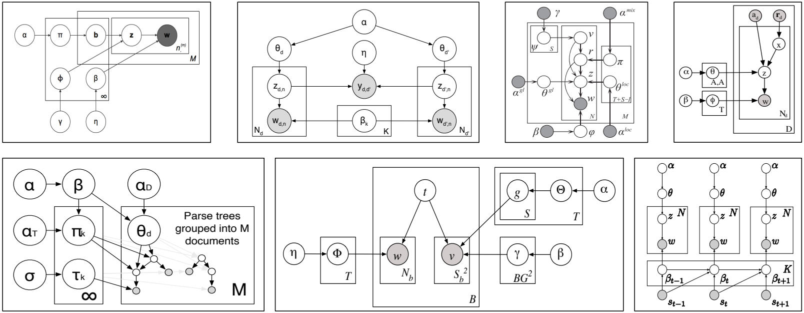


- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.

- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.

- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999)
  It is a mixed-membership model (Erosheva, 2004).
  It relates to PCA and matrix factorization (Jakulin and Buntine, 2002).
  Was independently invented for genetics (Pritchard et al., 2000)

LDA summary

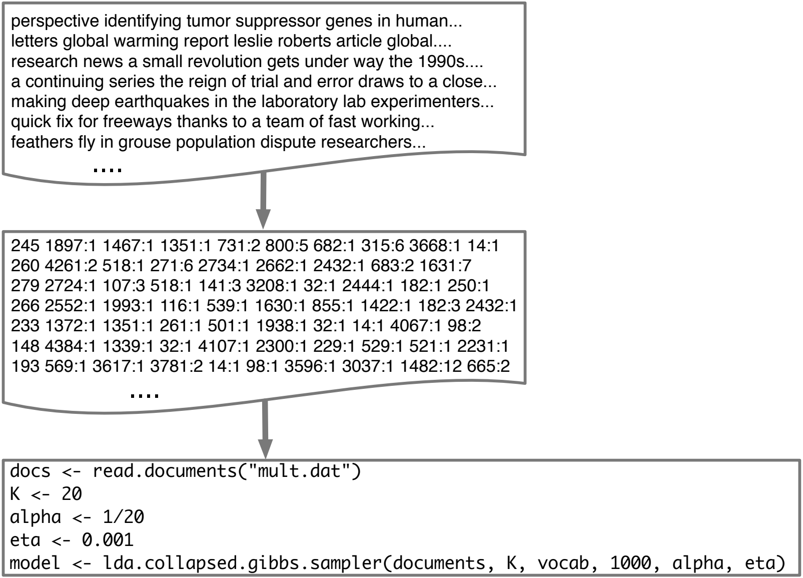


- LDA is a simple building block that enables many applications.

- It is popular because organizing and finding patterns in data has become important in the sciences, humanties, industry, and culture.

- Further, algorithmic improvements let us fit models to massive data.

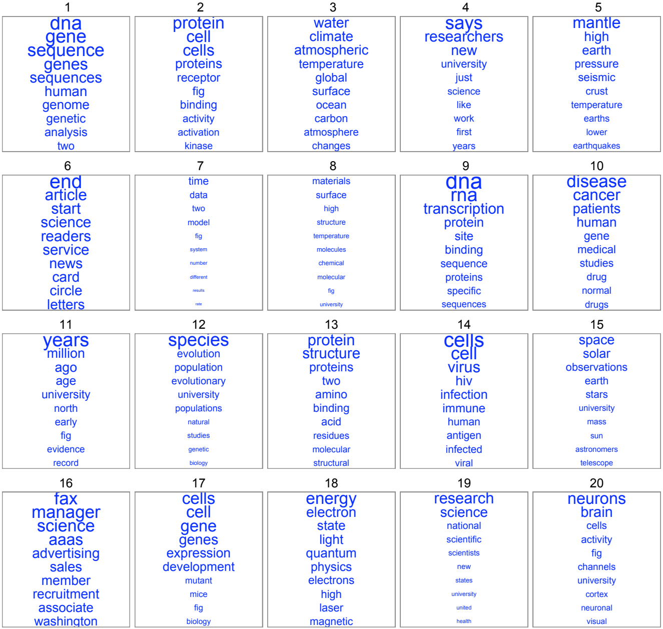
Example: LDA in R (Jonathan Chang)



perspective identifying tumor suppressor genes in human...
letters global warming report leslie roberts article global....
research news a small revolution gets under way the 1990s....
a continuing series the reign of trial and error draws to a close...
making deep earthquakes in the laboratory lab experimenters...
quick fix for freeways thanks to a team of fast working...
feathers fly in grouse population dispute researchers...
....

```
245 1897:1 1467:1 1351:1 731:2 800:5 682:1 315:6 3668:1 14:1
260 4261:2 518:1 271:6 2734:1 2662:1 2432:1 683:2 1631:7
279 2724:1 107:3 518:1 141:3 3208:1 32:1 2444:1 182:1 250:1
266 2552:1 1993:1 116:1 539:1 1630:1 855:1 1422:1 182:3 2432:1
233 1372:1 1351:1 261:1 501:1 1938:1 32:1 14:1 4067:1 98:2
148 4384:1 1339:1 32:1 4107:1 2300:1 229:1 529:1 521:1 2231:1
193 569:1 3617:1 3781:2 14:1 98:1 3596:1 3037:1 1482:12 665:2
....
```

```
docs <- read.documents("mult.dat")
K <- 20
alpha <- 1/20
eta <- 0.001
model <- lda.collapsed.gibbs.sampler(documents, K, vocab, 1000, alpha, eta)
```



1
dna
gene
sequence
genes
sequences
human
genome
genetic
analysis
two

2
protein
cell
cells
proteins
receptor
fig
binding
activity
activation
kinase

3
water
climate
atmospheric
temperature
global
surface
ocean
carbon
atmosphere
changes

4
says
researchers
new
university
just
science
like
work
first
years

5
mantle
high
earth
pressure
seismic
crust
temperature
earths
lower
earthquakes

6
end
article
start
science
readers
service
news
card
circle
letters

7
time
data
two
model
fig
system
number
different
results
rate

8
materials
surface
high
structure
temperature
molecules
chemical
molecular
fig
university

9
dna
rna
transcription
protein
site
binding
sequence
proteins
specific
sequences

10
disease
cancer
patients
human
gene
medical
studies
drug
normal
drugs

11
years
million
ago
age
university
north
early
fig
evidence
record

12
species
evolution
population
evolutionary
university
populations
natural
studies
genetic
biology

13
protein
structure
proteins
two
amino
binding
acid
residues
molecular
structural

14
cells
cell
virus
hiv
infection
immune
human
antigen
infected
viral

15
space
solar
observations
earth
stars
university
mass
sun
astronomers
telescope

16
fax
manager
science
aaas
advertising
sales
member
recruitment
associate
washington

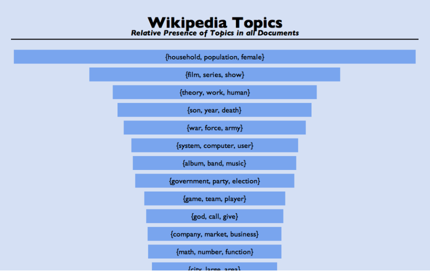
17
cells
cell
gene
genes
expression
development
mutant
mice
fig
biology

18
energy
electron
state
light
quantum
physics
electrons
high
laser
magnetic

19
research
science
national
scientific
scientists
new
states
university
united
health

20
neurons
brain
cells
activity
fig
channels
university
cortex
neuronal
visual

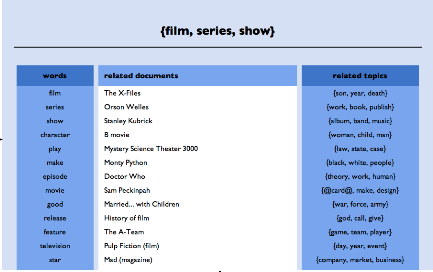
Open source document browser (with Allison Chaney)



Wikipedia Topics

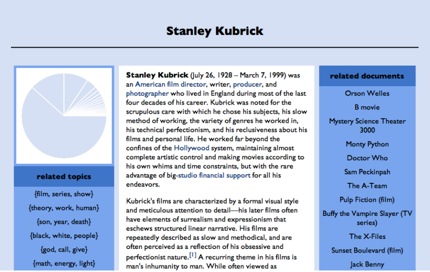
Relative Presence of Topics in all Documents

- {household, population, female}
- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {war, force, army}
- {system, computer, user}
- {album, band, music}
- {government, party, election}
- {game, team, player}
- {god, call, give}
- {company, market, business}
- {math, number, function}
- {city, large, area}



{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}



Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

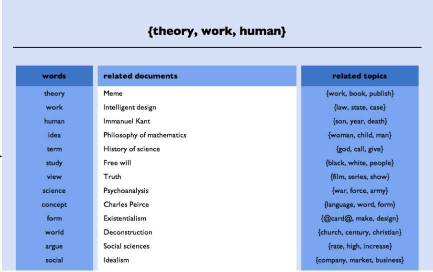
Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related topics

- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {black, white, people}
- {god, call, give}
- {math, energy, light}

related documents

- Orson Welles
- B movie
- Mystery Science Theater 3000
- Monty Python
- Doctor Who
- Sam Peckinpah
- The A-Team
- Pulp Fiction (film)
- Buffy the Vampire Slayer (TV series)
- The X-Files
- Sunset Boulevard (film)
- Jack Benny

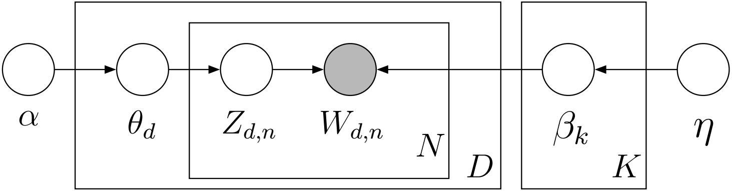


{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

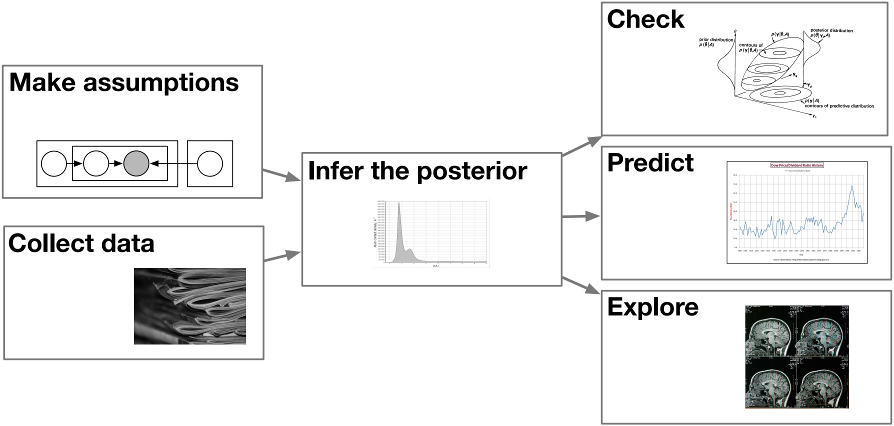
Beyond Latent Dirichlet Allocation

Extending LDA



- LDA is a simple topic model.
- It can be used to find topics that describe a corpus.
- Each document exhibits multiple topics.
- How can we build on this simple model of text?

Extending LDA



Make assumptions

Collect data

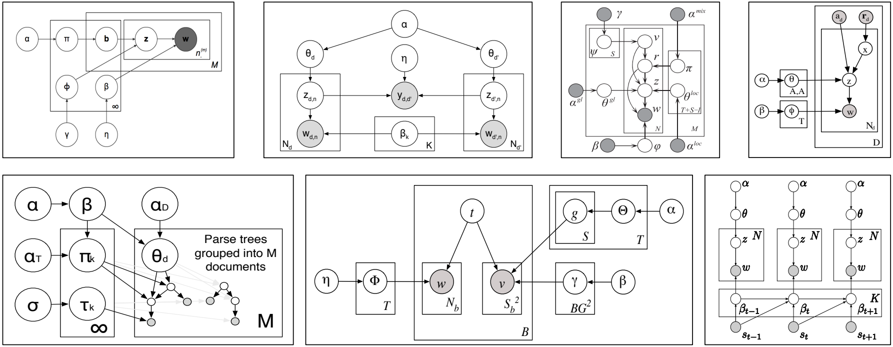
Infer the posterior

Check

Predict

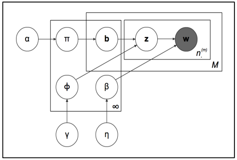
Explore

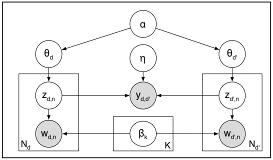
Extending LDA

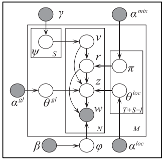


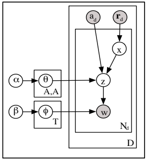
- LDA can be **embedded in more complicated models**, embodying further intuitions about the structure of the texts.
- E.g., it can be used in models that account for syntax, authorship, word sense, dynamics, correlation, hierarchies, and other structure.

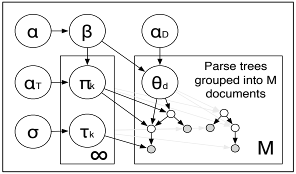
Extending LDA

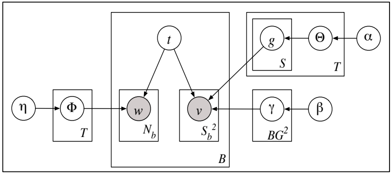


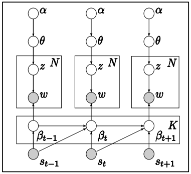








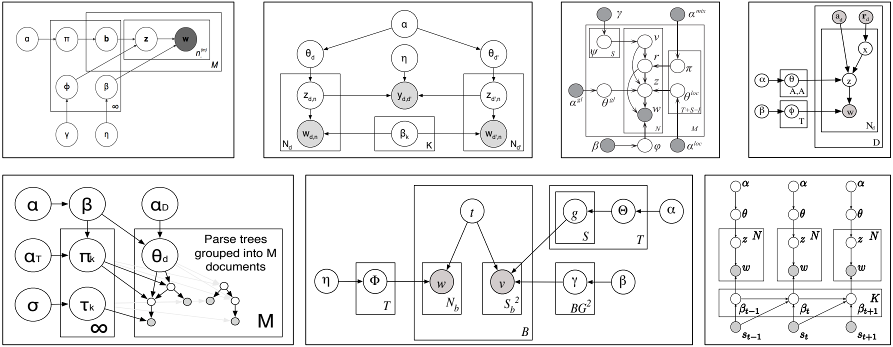




- The **data generating distribution** can be changed. We can apply mixed-membership assumptions to many kinds of data.

- E.g., we can build models of images, social networks, music, purchase histories, computer code, genetic data, and other types.

Extending LDA



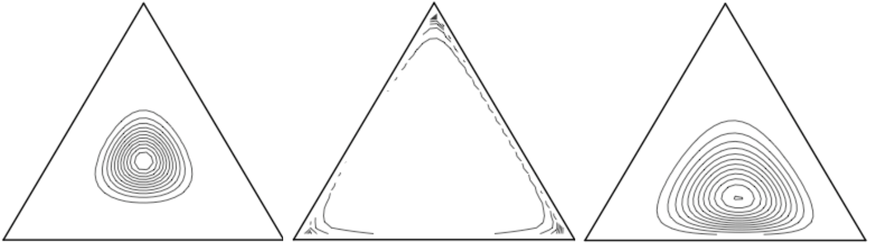
- The **posterior** can be used in creative ways.
- E.g., we can use inferences in information retrieval, recommendation, similarity, visualization, summarization, and other applications.

Extending LDA

- These different kinds of extensions can be combined.
- (Really, these ways of extending LDA are a big advantage of using **probabilistic modeling** to analyze data.)
- To give a sense of how LDA can be extended, I'll describe several examples of extensions that my group has worked on.
- We will discuss
  - **Correlated topic models**
  - **Dynamic topic models & measuring scholarly impact**
  - **Supervised topic models**
  - **Relational topic models**
  - **Ideal point topic models**
  - **Collaborative topic models**

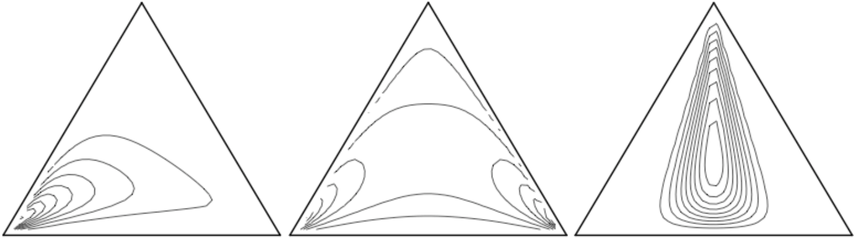
Correlated and Dynamic Topic Models

Correlated topic models



- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

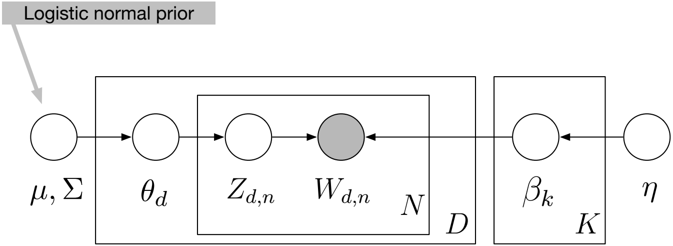
Correlated topic models



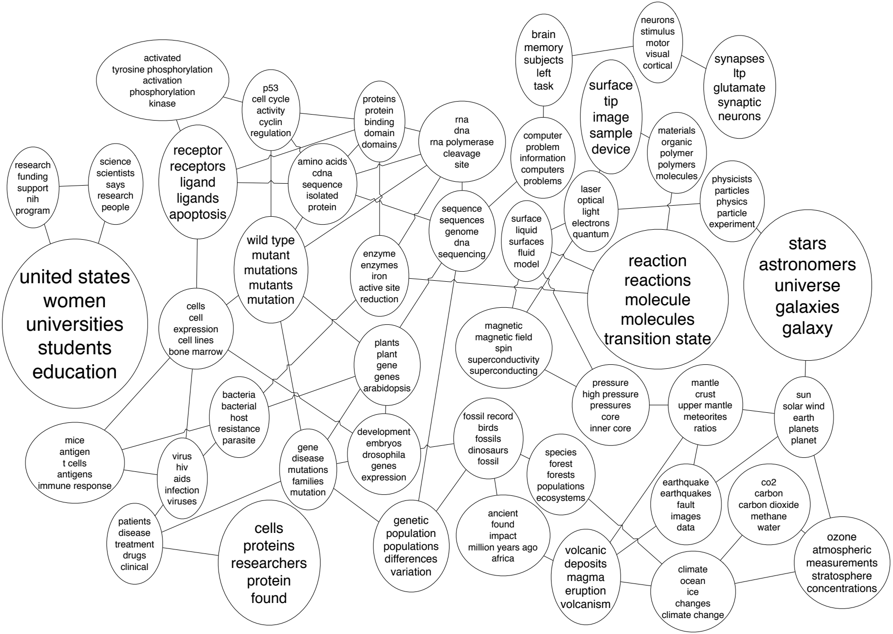
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$
\begin{aligned}
X &\sim \mathcal{N}_K(\mu, \Sigma) \\
\theta_i &\propto \exp\{x_i\}.
\end{aligned}
$$

Correlated topic models



- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but computation is more complex



activated
tyrosine phosphorylation
activation
phosphorylation
kinase
p53
cell cycle
activity
cyclin
regulation
proteins
protein
binding
domain
domains
brain
memory
subjects
left
task
neurons
stimulus
motor
visual
cortical
synapses
ltp
glutamate
synaptic
neurons
surface
tip
image
sample
device
research
funding
support
nih
program
science
scientists
says
research
people
receptor
receptors
ligand
ligands
apoptosis
amino acids
cdna
sequence
isolated
protein
rna
dna
rna polymerase
cleavage
site
computer
problem
information
computers
problems
materials
organic
polymer
polymers
molecules
laser
optical
light
electrons
quantum
physicists
particles
physics
particle
experiment
sequence
sequences
genome
dna
sequencing
surface
liquid
surfaces
fluid
model
wild type
mutant
mutations
mutants
mutation
enzyme
enzymes
iron
active site
reduction
united states
women
universities
students
education
reaction
reactions
molecule
molecules
transition state
stars
astronomers
universe
galaxies
galaxy
cells
cell
expression
cell lines
bone marrow
magnetic
magnetic field
spin
superconductivity
superconducting
plants
plant
gene
genes
arabidopsis
pressure
high pressure
pressures
core
inner core
mantle
crust
upper mantle
meteorites
ratios
sun
solar wind
earth
planets
planet
bacteria
bacterial
host
resistance
parasite
fossil record
birds
fossils
dinosaurs
fossil
development
embryos
drosophila
genes
expression
mice
antigen
t cells
antigens
immune response
virus
hiv
aids
infection
viruses
gene
disease
mutations
families
mutation
species
forest
forests
populations
ecosystems
earthquake
earthquakes
fault
images
data
co2
carbon
carbon dioxide
methane
water
patients
disease
treatment
drugs
clinical
cells
proteins
researchers
protein
found
genetic
population
populations
differences
variation
ancient
found
impact
million years ago
africa
volcanic
deposits
magma
eruption
volcanism
climate
ocean
ice
changes
climate change
ozone
atmospheric
measurements
stratosphere
concentrations

Dynamic topic models

1789

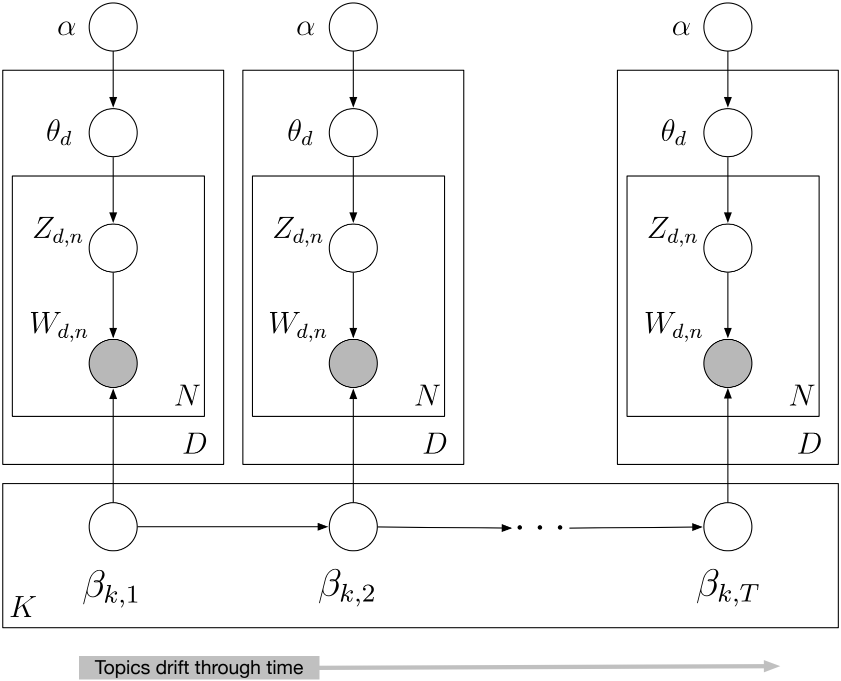
My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

Inaugural addresses

2009

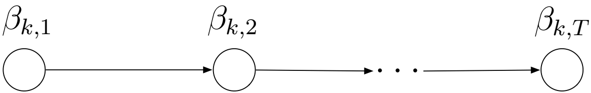
AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics *drift* in a sequence.



α
α
α
θ_d
θ_d
θ_d
$Z_{d,n}$
$Z_{d,n}$
$Z_{d,n}$
$W_{d,n}$
$W_{d,n}$
$W_{d,n}$
N
N
N
D
D
D
$\beta_{k,1}$
$\beta_{k,2}$
$\beta_{k,T}$
K
Topics drift through time

Dynamic topic models



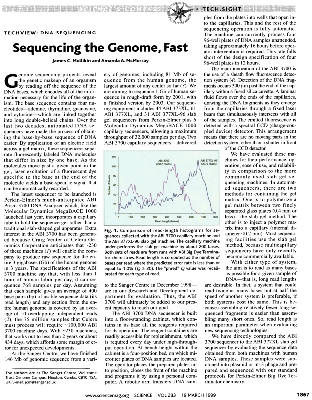
- Use a logistic normal distribution to model topics evolving over time.
- Embed it in a state-space model on the log of the topic distribution

$$
\begin{aligned}
\beta_{t,k} \,|\, \beta_{t-1,k} &\sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2) \\
p(w \,|\, \beta_{t,k}) &\propto \exp\{\beta_{t,k}\}
\end{aligned}
$$

- As for CTMs, this makes computation more complex. But it lets us make inferences about sequences of documents.

Dynamic topic models

Original article



TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 µm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

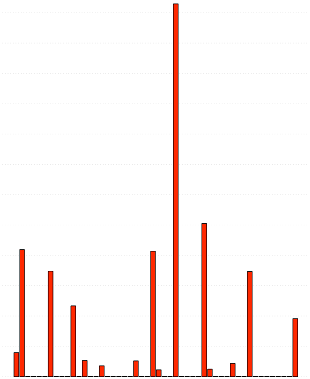
We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

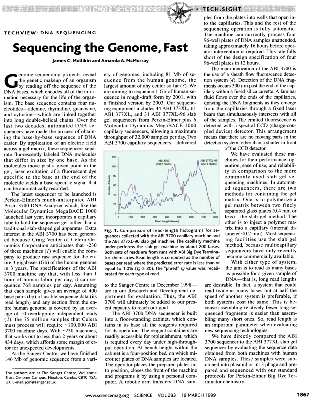
The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

Topic proportions



Dynamic topic models

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 µm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

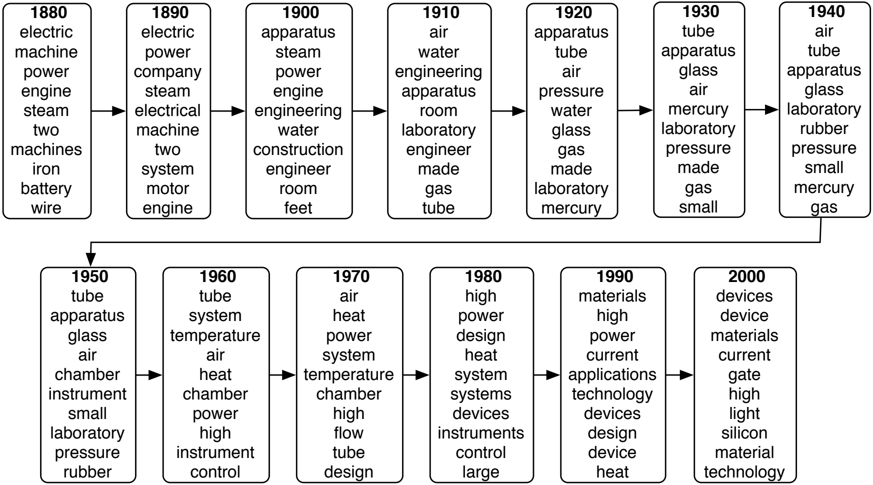
Most likely words from top topics

sequence
genome
genes
sequences
human
gene
dna
sequencing
chromosome
regions
analysis
data
genomic
number

devices
device
materials
current
high
gate
light
silicon
material
technology
electrical
fiber
power
based

data
information
network
web
computer
language
networks
time
software
system
words
algorithm
number
internet

Dynamic topic models



1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire

1890
electric
power
company
steam
electrical
machine
two
system
motor
engine

1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet

1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube

1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury

1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small

1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas

1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber

1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control

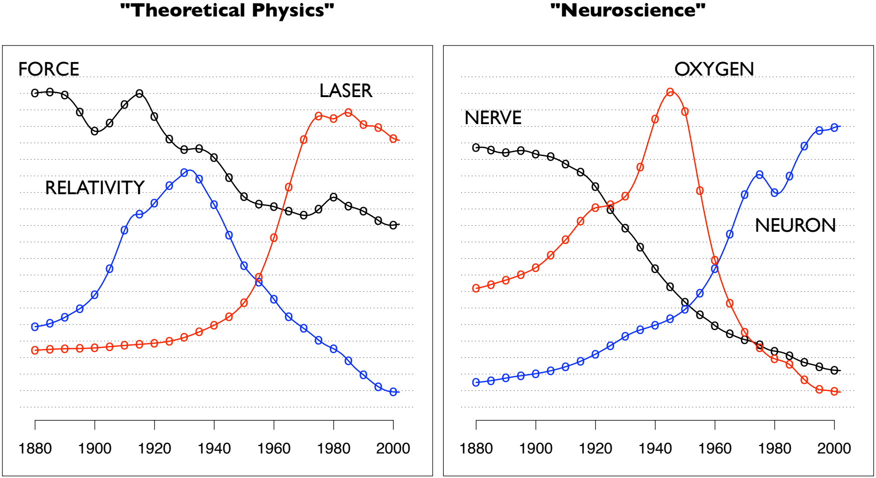
1970
air
heat
power
system
temperature
chamber
high
flow
tube
design

1980
high
power
design
heat
system
systems
devices
instruments
control
large

1990
materials
high
power
current
applications
technology
devices
design
device
heat

2000
devices
device
materials
current
gate
high
light
silicon
material
technology

Dynamic topic models



Dynamic topic models

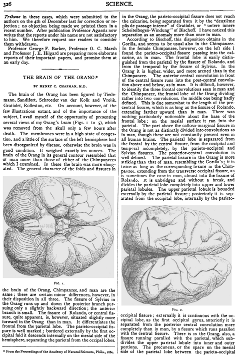
- **Time-corrected similarity** shows a new way of using the posterior.
- Consider the expected Hellinger distance between the topic proportions of two documents,

$$d_{ij} = \mathrm{E}\left[\sum_{k=1}^{K}(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \,|\, \mathbf{w}_i, \mathbf{w}_j\right]$$

- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year.
- Similarity based only on topic proportions

Dynamic topic models

The Brain of the Orang (1880)



326 SCIENCE.

Tribune in these cases, which were submitted to the authors on the 4th of December last for correction or rejection; no objection being made we printed them in a recent number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory to him. We therefore request our readers to consider them withdrawn.

Professor George F. Barker, Professor O. C. Marsh and Professor J. E. Hilgard are preparing more elaborate reports of their important papers, and promise them at an early day.

THE BRAIN OF THE ORANG.*

BY HENRY C. CHAPMAN, M.D.

The brain of the Orang has been figured by Tiedemann, Sandifort, Schroeder van der Kolk and Vrolik, Gratiolet, Rolleston, etc. On account, however, of the few illustrations extant, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my Orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours after death. The membranes were in a high state of congestion, and a little of the surface of the left hemisphere had been disorganized by disease, otherwise the brain was in good condition. It weighed exactly ten ounces. The brain of the Orang in its general contour resembled that of man more than those of either of the Chimpanzees which I examined. In these the brain was more elongated. The general character of the folds and fissures in the brain of the Orang, Chimpanzee, and man are the same; there are certain minor differences, however, in their disposition in all three. The fissure of Sylvius in the Orang runs up and down the posterior branch pursuing only a slightly backward direction; the anterior branch is small. The fissure of Rolando, or central fissure, quite apparent, is, however, situated slightly more forward in the Orang than in man. It differentiates the frontal from the parietal lobe. The parieto-occipital fispure is well marked; bordered externally by the first occipital fold it descends internally on the mesial side of the hemisphere, separating the parietal from the occipal lobes.

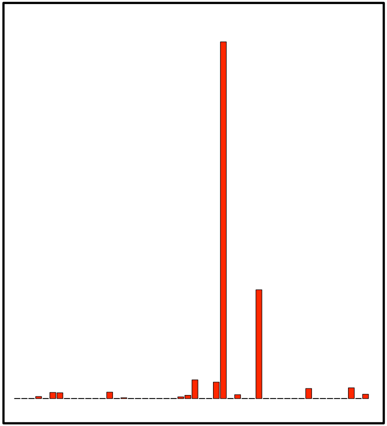
FIG. 1.

in the Orang, the parieto-occipital fissure does not reach the calcarine, being separated from it by the "deuxième plis de passage interne" of Gratiolet, or "untere innere Scheitelbogen-Windung" of Bischoff. I have noticed this separation as an anomaly more than once in man.

According to Bischoff, this disposition obtains in the Gorilla, and seems to be usual also in the Chimpanzee. In the female Chimpanzee, however, on the left side I found the parieto-occipital fissure passing into the calcarine, as in man. The frontal lobe is easily distinguished from the parietal by the fissure of Rolando, and from the temporal by the fissure of Sylvius. In the Orang it is higher, wider, and more arched than in the Chimpanzee. The anterior central convolution in front of the central fissure runs into the post-central convolution above and below, as in man. It is difficult, however, to identify the three frontal convolutions seen in man and the Chimpanzee, the frontal lobe of the Orang dividing rather into two convolutions, the middle one being badly defined. This is due somewhat to the length of the precentral fissure, which is as long as the fissure of Rolando, extending farther upward than in man. There was nothing particularly noticeable about the base of the frontal lobe; on the mesial surface it ran into the parietal. The part above the calloso-marginal fissure in the Orang is not as distinctly divided into convolutions as in man, though these are not constantly present even in all human brains. The parietal lobe is separated from the frontal by the central fissure, from the occipital and temporal incompletely, by the parieto-occipital and Sylvian fissures. The posterior-central convolution is well defined. The parietal fissure in the Orang is more striking than that of man, resembling the Gorilla's; it is twice as long as the corresponding fissure in the Chimpanzee, extending from the transverse occipital fissure, as is sometimes the case in man, almost into the fissure of Rolando. It is unbridged and without a break, and divides the parietal lobe completely into upper and lower parietal lobules. The upper parietal lobule is bounded externally by the parietal fissure; posteriorly it is separated from the occipital lobe, internally by the parieto-occipital fissure; externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution more completely than in man, by a fissure which runs paralled with the central fissure. There is in the Orang, also, a fissure running paralled with the parietal, which subdivides the upper parietal lobule into inner and outer portions. The precuneus, or the space on the mesial side of the parietal lobe between the parieto-occipital

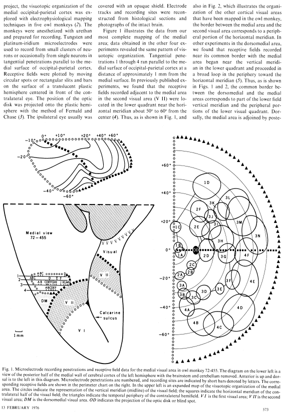
FIG. 2.

[*] From the Proceedings of the Academy of Natural Sciences, Phila., 1880.



Dynamic topic models

Representation of the Visual Field on the Medial Wall
of Occipital-Parietal Cortex in the Owl Monkey (1976)

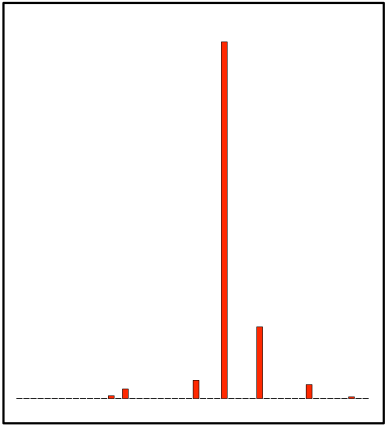


project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (2). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from small clusters of neurons or occasionally from single neurons in tangential penetrations parallel to the medial surface of occipital-parietal cortex. Receptive fields were plotted by moving circular spots or rectangular slits and bars on the surface of a translucent plastic hemisphere centered in front of the contralateral eye. The position of the optic disk was projected onto the plastic hemisphere with the method of Fernald and Chase (3). The ipsilateral eye usually was covered with an opaque shield. Electrode tracks and recording sites were reconstructed from histological sections and photographs of the intact brain.

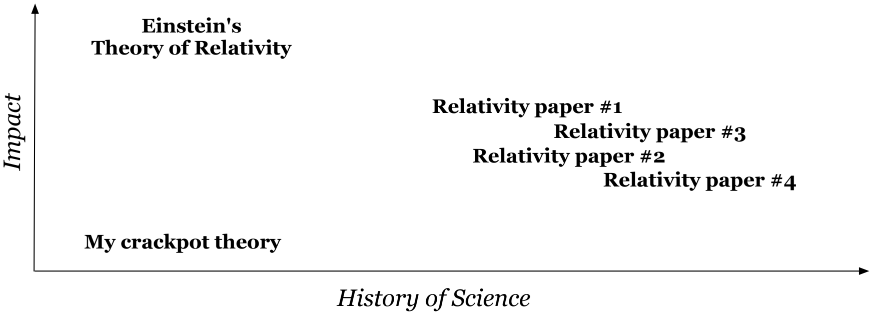
Figure 1 illustrates the data from our most complete mapping of the medial area; data obtained in the other four experiments revealed the same pattern of visuotopic organization. Tangential penetrations 1 through 4 ran parallel to the medial surface of occipital-parietal cortex at a distance of approximately 1 mm from the medial surface. In previously published experiments, we found that the receptive fields recorded adjacent to the medial area in the second visual area (V II) were located in the lower quadrant near the horizontal meridian about 50° to 60° from the center (4). Thus, as is shown in Fig. 1, and also in Fig. 2, which illustrates the organization of the other cortical visual areas that have been mapped in the owl monkey, the border between the medial area and the second visual area corresponds to a peripheral portion of the horizontal meridian. In other experiments in the dorsomedial area, we found that receptive fields recorded near its common border with the medial area began near the vertical meridian in the lower quadrant and proceeded in a broad loop in the periphery toward the horizontal meridian (5). Thus, as is shown in Figs. 1 and 2, the common border between the dorsomedial and the medial areas corresponds to part of the lower field vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjoined by poste-

Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the left in this diagram. Microelectrode penetrations are numbered, and recording sites are indicated by short bars denoted by letters. The corresponding receptive fields are shown in the perimeter chart on the right. In the upper left is an expanded map of the visuotopic organization of the medial area. The circles indicate the representation of the vertical meridian (midline) of the visual field; the squares indicate the horizontal meridian of the contralateral half of the visual field; the triangles indicate the temporal periphery of the contralateral hemifield. *V I* is the first visual area; *V II* is the second visual area; *DM* is the dorsomedial visual area. *OD* indicates the projection of the optic disk or blind spot.

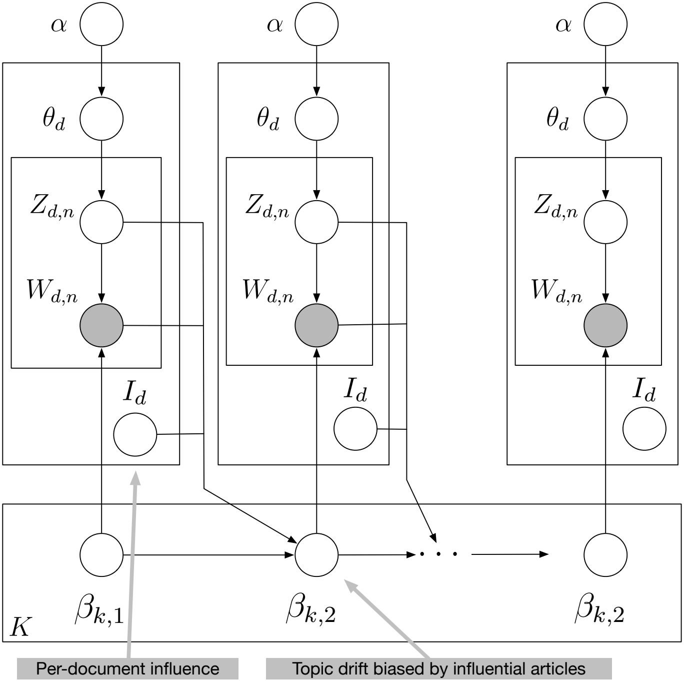
13 FEBRUARY 1976 573



Measuring scholarly impact



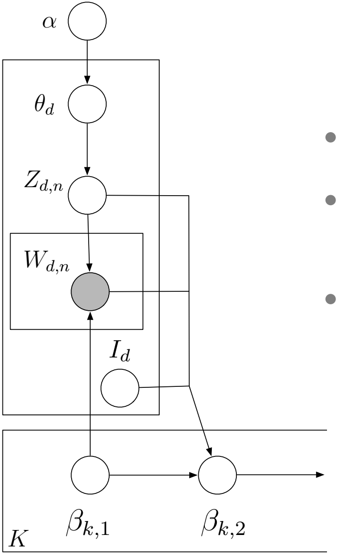
- We built on the DTM to measure **scholarly impact** with sequences of text.
- Influential articles reflect future changes in language use.
- The “influence” of an article is a latent variable.
- Influential articles affect the drift of the topics that they discuss.
- The posterior gives a retrospective estimate of influential articles.



Per-document influence

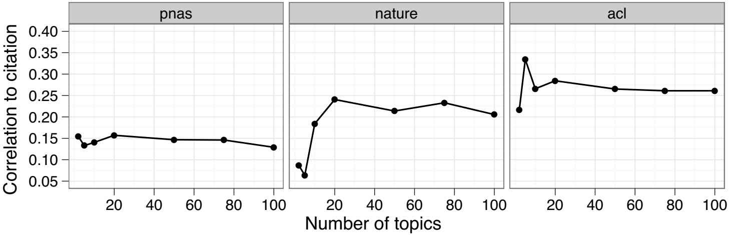
Topic drift biased by influential articles

Measuring scholarly impact



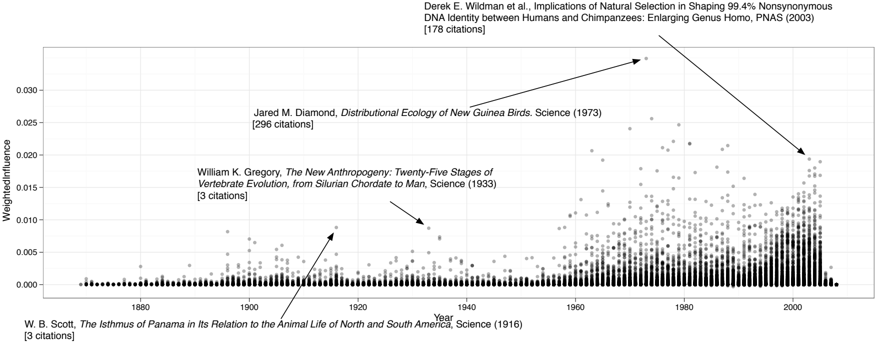
- Each document has an influence score I_d.
- Each topic drifts in a way that is biased towards the documents with high influence.
- We can examine the posterior of the influence scores to retrospectively find articles that best explain the changes in language.

Measuring scholarly impact



- This measure of impact only uses the words of the documents.
  It correlates strongly with citation counts.
- High impact, high citation: “The Mathematics of Statistical Machine Translation: Parameter Estimation” (Brown et al., 1993)
- “Low” impact, high citation: “Building a large annotated corpus of English: the Penn Treebank” (Marcus et al., 1993)

Measuring scholarly impact



Derek E. Wildman et al., Implications of Natural Selection in Shaping 99.4% Nonsynonymous DNA Identity between Humans and Chimpanzees: Enlarging Genus Homo, PNAS (2003) [178 citations]

Jared M. Diamond, *Distributional Ecology of New Guinea Birds.* Science (1973) [296 citations]

William K. Gregory, *The New Anthropogeny: Twenty-Five Stages of Vertebrate Evolution, from Silurian Chordate to Man*, Science (1933) [3 citations]

W. B. Scott, *The Isthmus of Panama in Its Relation to the Animal Life of North and South America*, Science (1916) [3 citations]

- PNAS, *Science*, and *Nature* from 1880–2005
- 350,000 Articles
- 163M observations
- Year-corrected correlation is 0.166

Summary: Correlated and dynamic topic models

- The Dirichlet assumption on topics and topic proportions makes strong conditional independence assumptions about the data.
- The **correlated topic model** uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur.
- The **dynamic topic model** uses a logistic normal in a linear dynamic model to capture how topics change over time.
- What’s the catch? These models are harder to compute with. (Stay tuned.)

Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey

Hamed Jelodar[1], Yongli Wang[1], Chi Yuan[1], Xia Feng[2]
Department of Computer Science and Engineering
Nanjing University of Science and Technology,
Nanjing - 210094, China
{Jelodar, Yongliwang, Yuanchi}@njust.edu.cn[1], 779477284@qq.com[2]

Abstract:
Topic modeling is one of the most powerful techniques in text mining for data mining, latent data discovery, and finding relationships among data, text documents. Researchers have published many articles in the field of topic modeling and applied in various fields such as software engineering, political science, medical and linguistic science, etc. There are various methods for topic modeling, which Latent Dirichlet allocation (LDA) is one of the most popular methods in this field. Researchers have proposed various models based on the LDA in topic modeling. According to previous work, this paper can be very useful and valuable for introducing LDA approaches in topic modeling. In this paper, we investigated scholarly articles highly (between 2003 to 2016) related to Topic Modeling based on LDA to discover the research development, current trends and intellectual structure of topic modeling. Also, we summarize challenges and introduce famous tools and datasets in topic modeling based on LDA.

Keywords: Topic modelling, Gibbs Sampling, Latent Dirichlet Allocation, Expectation Maximization, LDA

1. Introduction

Topic models (TM) are a well-know and significant modern machine learning technology that has been widely used in text mining, network analysis and genetics, and more other domains. Topic models are prominent for demonstrating discrete data; also, give a productive approach to find hidden structures/semantics in gigantic information. There are many papers for in this field and definitely cannot mention to all of them, so we selected more signification papers. Topic models are applied in various fields including medical sciences (Zhang et al., 2017) (Jiang et al., 2012) (Paul and Dredze, 2011) (Wu et al., 2012b) , software engineering (Linstead et al., 2007) (Gethers and Poshyvanyk, 2010) (Asuncion et al., 2010) (Thomas, 2011) (Thomas et al., 2011), geography (Cristani et al., 2008) (Eisenstein et al., 2010) (Tang et al., 2013) (Yin et al., 2011) (Sizov, 2010), political science (Chen et al., 2010) (Cohen and Ruths, 2013) (Greene and Cross, 2015) .

From an applied perspective in the field of political science, **Greene et al.** proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time and identify both niche topics related to events at a particular point in time and broad, long-running topics. This paper has focused on European Parliament speeches, the proposed topic modeling method has a number of potential applications in the study of politics, including the analysis of speeches in other parliaments, political manifestos, and other more traditional forms of political texts (Greene and Cross, 2015). Other researchers have also proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time and identify both niche topics related to events at a particular point in time and broad, long-running topics. This paper has focused on European Parliament speeches; the proposed topic modeling method has a number of potential applications in the study of politics, including the analysis of speeches in other parliaments, political manifestos, and other more traditional forms of political texts [17]. **Fang et al.**

suggested a new unsupervised topic model based on LDA for contrastive opinion modeling which purpose to find the opinions from multiple views, according to a given topic and their difference on the topic with qualifying criteria, the model called Cross-Perspective Topic (CPT) model. They performed experiments with both qualitative and quantitative measures on two datasets in the political area that include: first dataset is statement records of U.S. senators that show political stances of senators by these records, also for the second dataset, extracted of world News Medias from three representative media in U.S (New York Times), China (Xinhua News) and India (Hindu). To evaluate their approach with other models used corrIDA and LDA as two baselines (Fang et al., 2012).

Another group of researchers focused on topic modeling in software Engineering, **Linstead et al.** For the first time, they used LDA, to extract topics in source code and perform to visualization of software similarity, In other words, LDA use an intuitive approach for calculation of similarity between source files with obtain their respective distributions of each document over topics. They utilized their method on 1,555 software projects from Apache and SourceForge that includes 19 million source lines of code (SLOC). The authors demonstrated this approach, can be effective for project organization, software refactoring (Linstead et al., 2007). **Tian et al.** introduced a method based on LDA for automatically categorizing software systems, called LACT. For evaluation of LACT, used 43 open-source software systems in different programming languages and showed LACT can categorization of software systems based on type of programming language (Tian et al., 2009). **Lukinet al.** Proposed an approach topic modeling based on LDA model for the purpose of bug localization. Their idea, applied to analysis of same bugs in Mozilla and Eclipse and result showed that their LDA-based approach is better than LSI for evaluate and analyze of bugs in these source codes (Lukins et al., 2008, Lukins et al., 2010).

An analysis of geographic information is another issue that can be referred to Sizov **et al.** They introduced a novel method based on multi-modal Bayesian models to describe social media by merging text features and spatial knowledge that called GeoFolk. As a general outlook, this method can be considered as an extension of Latent Dirichlet Allocation (LDA). They used the available standard CoPhIR dataset that it contains an abundance of over 54 million Flickr. The GeoFolk model has the ability to be used in quality-oriented applications and can be merged with some models from Web 2.0 social (Sizov, 2010)**. Yin et al.** This article examines the issue of topic modeling to extract the topics from geographic information and GPS-related documents. They proposed three strategies of modeling geographical topics including , text-driven model, location-driven model and a novel joint model called LGTA (Latent Geographical Topic Analysis) that is a combination of topic modeling and geographical clustering. To test their approaches, they collected a set of data from the website Flickr, according to various topics (Yin et al., 2011).

In other view, According to our knowledge, most of the previous studies had various goals, such as: Source code analysis (Linstead et al., 2007) (Lukins et al., 2010) (Linstead et al., 2008) (Tian et al., 2009) (Chen et al., 2012) (Gethers and Poshyvanyk, 2010) (Savage et al., 2010), Opinion and aspect Mining (Chen et al., 2010) (Zheng et al., 2014) (Cheng et al., 2014) (Zhai et al., 2011) (Bagheri et al., 2014) (Wang et al., 2014c) (Xianghua et al., 2013, Jo and Oh, 2011) (Paul and Girju, 2010) (Titov and McDonald, 2008), Event detection (Qian et al., 2016) (Hu et al., 2012, Weng and Lee, 2011) (Lin et al., 2010), image classification (Cristani et al., 2008) (Wang and Mori, 2011), system recommendation (Zoghbi et al., 2016) (Cheng and Shen, 2016) (Zhao et al., 2016) (Lu and Lee, 2015) (Wang et al., 2014a) (Yang and Rim, 2014) (Kim and Shim, 2014) and emotion classification(Roberts et al., 2012) (Rao,

2016) (Rao et al., 2014), etc. For example in gforecommendation system, **Zhao and et al.** proposed a personalized hashtag recommendation approach based LDA model that can discover latent topics in microblogs, called Hashtag-LDA and applied experiments on 'UDI-TwitterCrawl-Aug2012-Tweets' as a real-world Twitter dataset(Zhao et al., 2016). **Jin and et al.** The authors focused on the issue of tag recommendation. They proposed hybrids approach based on a combination of Language Model (LM) and LDA for tag recommendation in terms of topic knowledge. The authors used a subset from Bibsonomy datset that including; 14,443 resources, 33,256 words 1,185 users, 13,276 tags, and 262,445 bookmarks in total. Finally, they found that combination of keyword and topic layer based approaches can be significantly effective to recommend new tags.

The main goal of this work is to provide an overview of the methods of topic modeling based on LDA. In summary, this paper makes four main contributions:

- We investigate scholarly articles (from 2003 to 2016) which are related to Topic Modeling based on LDA to discover the research development, current trends and intellectual structure of topic modeling based on LDA.
- We investigate topic modeling applications in various sciences.
- We summarize challenges in topic modeling, such as image processing, Visualizing topic models, Group discovery, User Behavior Modeling, and etc.
- We introduce some of the most famous data and tools in topic modeling.

2. Computer science and topic modeling

Topic models have an important role in computer science for text mining. In Topic modeling, a topic is a list of words that occur in statistically significant methods. A text can be an email, a book chapter, a blog posts, a journal article and any kind of unstructured text. Topic models cannot understand the means and concepts of words in text documents for topic modeling. Instead, they suppose that any part of the text is combined by selecting words from probable baskets of words where each basket corresponds to a topic. The tool goes via this process over and over again until it stays on the most probable distribution of words into baskets which call topics. Topic modeling can provide a useful view of a large collection in terms of the collection as a whole, the individual documents, and the relationships between the documents.

2.3 Latent Dirichlet Allocation

LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. Latent Dirichlet allocation (LDA), first introduced by Blei, Ng and Jordan in 2003(Blei et al., 2003), is one of the most popular methods in topic modeling. LDA represents topics by word probabilities. The words with highest probabilities in each topic usually give a good idea of what the topic is can word probabilities from LDA.

LDA, an unsupervised generative probabilistic method for modeling a corpus, is the most commonly used topic modeling method. LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well. Given a corpus D consisting of M documents, with document d having N d words ($d \in \{1,..., \mathrm{M}\}$), LDA models D according to the following generative process [4]:

*(a)*Choose a multinomial distribution φ_t for topic t ($t \in \{1,..., T\}$) from a Dirichlet distribution with parameter β.

(b) Choose a multinomial distribution θ_d for document d ($d \in \{1,..., M\}$) from a Dirichlet distribution with parameter α.

*(c)*For a word w_n ($n \in \{1,..., N_d\}$) in document d,

(i) Select a topic z_n from θ_d.

(ii) Select a word w_n from φ_{zn}.

In above generative process, words in documents are the only observed variables while others are latent variables (φ and θ) and hyper parameters (α and β). In order to infer the latent variables and hyper parameters, the probability of observed data D is computed and maximized as follows:

$$p(\mathrm{D}|\alpha\,,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\,\left(\sum_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{d_n}|z_{d_n},\varphi)P(\varphi|\beta)\right) d\theta_d d_p \qquad (1)$$

LDA is a distinguished tool for latent topic distribution for a large corpus. Therefore, it has the ability to identify sub-topics for a technology area composed of many patents, and represent each of the patents in an array of topic distributions. With LDA, the terms in the collection of documents produce a vocabulary that is then used to generate the latent topics. Documents are treated as a mixture of topics, where a topic is a probability distribution over this set of terms. Each document is then seen as a probability distribution over the set of topics. We can think of the data as coming from a generative process that is defined by the joint probability distribution over what is observed and what is hidden.

2.4 Parameter estimation, Inference, Training for LDA

Various methods have been proposed to estimate LDA parameters, such as variational method(Blei et al., 2003), expectation propagation(Minka and Lafferty, 2002) and Gibbs sampling(Griffiths and Steyvers, 2004).

- **Gibbs sampling** is a Monte Carlo Markov-chain algorithm, powerful technique in statistical inference**,** and a method of generating a sample from a joint distribution when only conditional distributions of each variable can be efficiently computed. According to our knowledge, researchers have widely used this method for the LDA. Some of works related based on LDA and Gibbs, such as (Xie et al., 2016) (Lu et al., 2016) (Yeh et al., 2016) (Rao, 2016) (Miao et al., 2016) (Panichella et al., 2013, Zhao et al., 2011) (Jagarlamudi and Daumé III, 2010) (Tian et al., 2009) (Ramage et al., 2009).
- **Expectation-Maximization (EM) algorithm** is a powerful method to obtain parameter estimation of graphical models and can use for unsupervised learning. In fact, the algorithm relies on discovering the maximum likelihood estimates of parameters when the data model depends on certain latent variables EM algorithm contains two steps, the E-step (expectation) and the M-step (maximization). Some researchers have applied this model to LDA training, such as (Zhu et al., 2009) (Guo et al., 2009) (Chang and Blei, 2009) (Blei and Jordan, 2003).
- **Variational Bayes inference** (VB), VB can be considered as a type of EM extension that uses a parametric approximation to the posterior distribution of both

parameters and other latent variables and attempts to optimize the fit (e.g. using KL-divergence) to the observed data. Some researchers have applied this model to LDA training, such as (Zhai et al., 2012) (Asuncion et al., 2010) (Chien and Chueh, 2011).

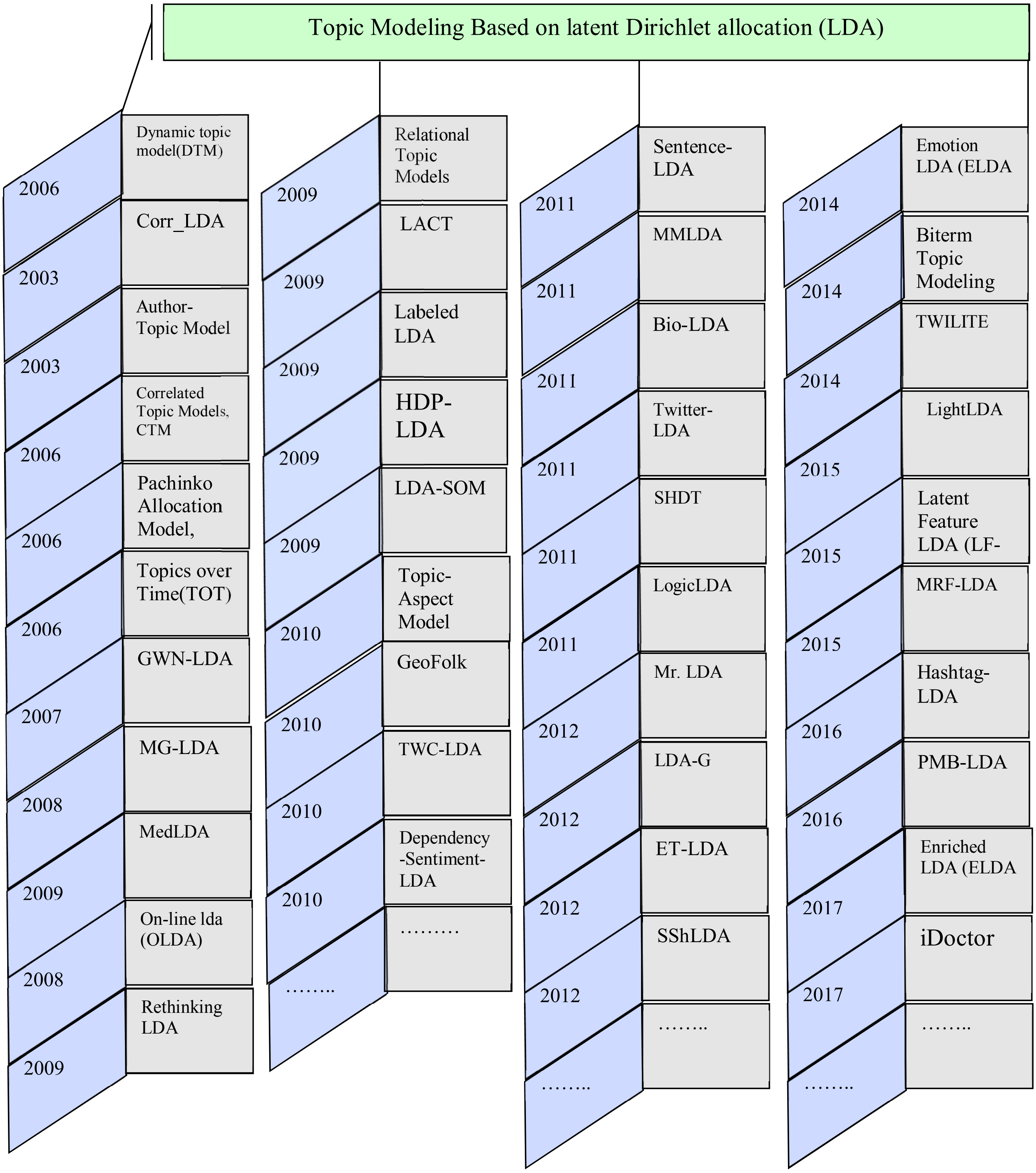


Fig1. Taxonomy of methods based on extension LDA, considered some of the impressive works

2.2.1. A brief look at past work: Research between 2003 to 2009

The LDA was first presented in 2003, and researchers have been tried to provide extended approaches based on LDA. Undeniably, this period (2003 to 2009) is very important because key and baseline approaches were introduced, such as: Corr_LDA, Author-Topic Model , DTM and , RTM etc .

DTM, Dynamic Topic Model (DTM) is introduced by Blei and Lafferty as an extension of LDA that this model can obtain the evolution of topics over time in a sequentially arranged corpus of documents and exhibits the evolution of word-topic distribution which causes it easy to vision the topic trend(Blei and Lafferty, 2006). **Lable LDA,** Labeled LDA (LLDA) is another of LDA extension which suppose that each document has a set of known labels (Ramage et al., 2009). This model can be trained with labeled documents and even supports documents with more than one label. Topics are learned from the co-occurring terms in places from the same category, with topics approximately capturing different place categories. A separate L-LDA model is trained for each place category, and can be used to infer the category of new, previously unseen places. LLDA is a supervised algorithm that makes topics applying the Labels assigned manually. Therefore, LLDA can obtain meaningful topics, with words that map well to the labels applied .

MedLDA, proposed the maximum entropy discrimination latent Dirichlet allocation (MedLDA) model, which incorporates the mechanism behind the hierarchical Bayesian models (such as, LDA) with the max margin learning (such as SVMs) according to a unified restricted optimization framework. In fact each data sample is considered to a point in a finite dimensional latent space, of which each feature corresponds to a topic, i.e., unigram distribution over terms in a vocabulary (Zhu et al., 2009). **Relational Topic Models (RTM),** is another extension, RTM is a hierarchical model of networks and per-node attribute data. First, each document was created from topics in LDA. Then, modelling the connections between documents and considered as binary variables, one for each pair from documents. These are distributed based on a distribution that depends on the topics used to generate each of the constituent documents. So in this way, the content of the documents are statistically linked to the link structure between them and we can say that this model can be used to summarize a network of documents (Chang and Blei, 2009).

Tablel 1. Some impressive articles based on LDA: between 2003- 2009

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Blei and Jordan, 2003)	Corr_LDA	2003	Variational EM	LDA	Image annotation and retrieval
(Rosen-Zvi et al., 2004)	Author-Topic Model	2004	Gibbs Sampling	LDA	Find the relationships between authors, documents, words, and topics
(McCallum et al., 2005)	AuthorRecipient-Topic (ART)	2005	Gibbs Sampling	-LDA -Author-Topic (AT)	Social network analysis and role discovery
(Blei and Lafferty, 2006)	Dynamic topic model(DTM)	2006	Kalman variational algorithm	LDA -Galton–Watson process	Provide a dynamic model for evolution of topics
(Wang and McCallum, 2006)	Topics over Time(TOT)	2006	Gibbs Sampling	LDA	Capture word co-occurrences and localization in continuous time.
(Li and McCallum, 2006)	Pachinko Allocation Model, PAM	2006	Gibbs Sampling	-LDA - a directed acyclic graph method	Capture arbitrary topic correlations
(Zhang et al., 2007)	GWN-LDA	2007	Gibbs sampling	LDA hierarchical Bayesian algorithm	Probabilistic community profile Discovery in social network
(AlSumait et al., 2008)	On-line lda (OLDA)	2008	Gibbs Sampling	-LDA -Empirical Bayes method	Tracking and Topic Detection
(Titov and McDonald, 2008)	MG-LDA	2008	Gibbs sampling	LDA	Sentiment analysis in multi-aspect
(Ramage et al., 2009)	Labeled LDA	2009	Gibbs Sampling	LDA	Producing a labeled document collection.
(Chang and Blei, 2009)	Relational Topic Models	2009	Expectation-maximization (EM)	LDA	Make predictions between nodes , attributes, links structure
(Wang and Blei, 2009)	HDP-LDA	2009	Gibbs Sampling	LDA	
(Lacoste-Julien et al., 2009)	DiscLDA	2009	Gibbs Sampling	LDA	Classification and

					dimensionality reduction in documents
(Tian et al., 2009)	LACT	2009	Gibbs sampling	LDA	Automatic Categorization of Software systems
(Wallach et al., 2009)	Rethinking LDA	2009	Gibbs Sampling	LDA	Data Discovery
(Guo et al., 2009)	WS-LDA	2009	Expectation-maximization (EM)	-LDA -query log	Query log mining
(Millar et al., 2009)	LDA-SOM	2009	Gibbs sampling	-LDA -self-organizing maps	Clustering and visualization of large documen
(Zhu et al., 2009)	MedLDA	2009	Expectation-maximization (EM)	-LDA -max-margin learning - support vector regression (SVR)	Regression and Classification

2.2.2. A brief look at past work: Research between 2010 to 2011

Eighteenth approaches are summarized in this subsection, where tenth are published in 2010 and Eighth in 2011. According to the Table 2, used LDA model for variety subjects, such as: Scientific topic discovery (Paul and Girju, 2010), Source code analysis (Savage et al., 2010), Opinion Mining (Zhai et al., 2011), Event detection (Lin et al., 2010), Image Classification (Wang and Mori, 2011) .

Sizov et al. introduced a novel method based on multi-modal Bayesian models to describe social media by merging text features and spatial knowledge that called GeoFolk. As a general outlook, this method can be considered as an extension of Latent Dirichlet Allocation (LDA). They used the available standard CoPhIR dataset that it contains an abundance of over 54 million Flickr. The GeoFolk model has the ability to be used in quality-oriented applications and can be merged with some models from Web 2.0 social (Sizov, 2010)**.**

Z. Zhai et al. use prior knowledge as a constraint in the LDA models to improve grouping of features by LDA. They extract must link and cannot-link constraint from the corpus. Must link indicates that two features must be in the same group while cannot-link restricts that two features cannot be in the same group. These constraints are extracted automatically. If at least one of the terms of two product features are same, they are considered to be in the same group as must link. On the other hand, if two features are expressed in the same sentence without conjunction ”and”, they are considered as a different feature and should be in different groups as cannot-link (Zhai et al., 2011).

Wang et al. they suggested an approach based on LDA that called Bio-LDA that can identify biological terminology to obtain latent topics. The authors have shown that this approach can be applied in different studies such as association search, association predication, and connectivity map generation. And they showed that Bio-LDA can be applied to increase the application of molecular bonding techniques as heat maps (Wang et al., 2011).

Tablel 2. Some impressive articles based on LDA: between 2010- 2011

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Sizov, 2010)	GeoFolk	2010	Gibbs sampling	LDA	Content management and retrieval of spatial information
(Jagarlamudi and Daumé III, 2010)	JointLDA	2010	Gibbs Sampling	-LDA -bag-of-word model	Mining multilingual topics
(Paul and Girju, 2010)	Topic-Aspect Model (TAM)	2010	Gibbs Sampling	-LDA -SVM	Scientific topic discovery
(Li et al., 2010)	Dependency-Sentiment-LDA	2010	Gibbs Sampling	LDA	Sentiment classification
(Savage et al., 2010)	TopicXP	2010		LDA	Source code analysis
(Zhai et al., 2011)	constrained-LDA	2010	Gibbs Sampling	LDA	Opinion Mining and Grouping Product Features
(Lin et al., 2010)	PET - Popular Events Tracking	2010	Gibbs Sampling	LDA	Event analysis in social network
(Weng and Lee, 2011) (Lin et al., 2010)	EDCoW	2010		-LDA - Wavelet Transformation	Event analysis in Twitter
(Wang et al., 2011)	Bio-LDA.	2011	Gibbs Sampling	-LDA	Extract biological terminology
(Zhao et al., 2011)	Twitter-LDA	2011	Gibbs Sampling	- LDA - author-topic model -PageRank	Extracting topical keyphrases and analyzing Twitter content
(Wang and Mori, 2011)	max-margin latent Dirichlet allocation (MMLDA	2011	variational inference	LDA - SVM	Image Classification and Annotation
(Jo and Oh, 2011)	Sentence-LDA	2011	Gibbs sampling	LDA	Aspects and sentiment discovery for web review
(Liu et al., 2011)	PLDA+	2011	Gibbs sampling	-LDA -weighted round-robin	Reduce inter-computer communication time
(Chien and Chueh, 2011)	Dirichlet class language model (DCLM),	2011	variational Bayesian EM (VB-EM) algorithm	speech recognition and exploitation of language models	Dirichlet class language model (DCLM),

2.2.3. A brief look at past work: Research between 2012 to 2013

According to Table 3, some of the popular works published between 2012 and 2013 focused on a variety of topics, such as music retrieve (Yoshii and Goto, 2012), opinion and aspect mining(Li et al., 2013), Event analysis(Hu et al., 2012).

ET-LDA, In this work, the authors developed a joint Bayesian model that performs event segmentation and topic modeling in one unified framework. In fact, they proposed an LDA model to obtain event's topics and analysis tweeting behaviors on Twitter that called Event and Tweets LDA (ET-LDA). They employed Gibbs Sampling method to estimate the topic distribution(Hu et al., 2012). **Mr. LDA ,** The authors introduced a novel model and parallelized LDA algorithm in the MapReduce framework that called Mr. LDA. In contrast other approaches which use Gibbs sampling for LDA, this model uses variational inference (Zhai et al., 2012). **LDA-GA,** The authors focused on the issue of Textual Analysis in Software Engineering. They proposed an LDA model based on Genetic Algorithm to determine a near-optimal configuration for LDA (LDA-GA), This approach is considered by three scenarios that include: (a) traceability link recovery, (b) feature location, and (c) labeling. They applied the Euclidean distance to measuring the distance between documents and used Fast collapsed Gibbs sampling to approximate the posterior distributions of parameters(Panichella et al., 2013).

Tablel 3. Some impressive articles based on LDA: between 2012- 2013

Author-study	**Model**	**Years**	**Parameter Estimation / Inference**	**Methods**	**Problem Domain**
(Wu et al., 2012a)	locally discriminative topic model (LDTM)	2012	Expectation-maximization (EM)	LDA	document semantic analysis
(Wu et al., 2012a)	locally discriminative topic model (LDTM)	2012	Expectation-maximization (EM)	LDA	document semantic analysis
(Hu et al., 2012)	ET-LDA	2012	Gibbs sampling	LDA	event segmentation Twitter
(Yoshii and Goto, 2012)	infinite latent harmonic allocation (iLHA)	2012	expectation–maximization (EM) algorithm /variational Bayes (VB)	-LDA -variational Bayes(VB) -HDP(Hierarchical Dirichlet processes)	multipitch analysis and music information retrieval
(Zhai et al., 2012)	Mr. LDA	2012	Variational Bayes inference	-LDA -Newton-Raphson method - MapReduce Algorithm	exploring document collections from large scale
(Tan et al., 2014)	FB-LDA , RCB-LDA	2012	Gibbs Sampling	LDA	analyze and track public sentiment variations (on

					twitter)
(Paul and Dredze, 2012)	factorial LDA	2012	Gibbs sampling	LDA	analysis text in a Multi-Dimensional multi-dimensional structure
(Mao et al., 2012)	SShLDA	2012	Gibbs sampling	LDA hLDA	Topic discovery in data space
(Choo et al., 2013)	Utopian	2013	Gibbs sampling	LDA	visual text analytics
(Panichella et al., 2013)	LDA–GA	2013	Gibbs sampling	LDA -Genetic Algorithm	software textual retrieval and analysis
(Xianghua et al., 2013)	Multi-aspect Sentiment Analysis for Chinese Online Social Reviews (MSA-COSRs	2013	Gibbs Sampling	LDA	sentiment analysis And aspect mining of of Chinese social reviews
(Li et al., 2013)	TopicSpam	2013	Gibbs sampling	LDA	opinion spam detection
(Chen et al., 2013)	WT-LDA	2013	Gibbs sampling	LDA	Web Service Clustering

2.2.4. A brief look at past work: Research between 2014 to 2015

According to Table 4, some of the popular works published between 2014 and 2015 focused on a variety of topics, such as: Hash/tag discovery (Wang et al., 2014a) (Lu and Lee, 2015), opinion mining and aspect mining (Bagheri et al., 2014) (Zheng et al., 2014) (Cheng et al., 2014) (Wang et al., 2014c), system recommendation(Lu and Lee, 2015) (Yang and Rim, 2014) (Kim and Shim, 2014).

Biterm Topic Modeling (BTM), Topic modeling over short texts is an increasingly important task due to the prevalence of short texts on the Web. Short texts are popular on today’s Web, especially with the emergence of social media. Inferring topics from large scale short texts becomes critical. They proposed a novel topic model for short texts, namely the biterm topic model (BTM). This model can well capture the topics within short texts by explicitly modeling word co-occurrence patterns in the whole corpus(Cohen et al., 2014).

TOT-MMM, introduced a hashtag recommendation that called TOT-MMM, This approach is a hybrid model that combines a temporal clustering component similar to that of the Topics-over-Time (TOT) Model with the Mixed Membership Model (MMM) that was originally proposed for word-citation co-occurrence. This model can capture the temporal clustering effect in latent topics, thereby improving hashtag modeling and recommendations. They developed a collapsed Gibbs sampling (CGS) for approximate the posterior modes of the remaining random variables(Lu and Lee, 2015). The posterior distribution of latent topic equaling ***k*** for the ***n***th hashtag in tweet ***d*** is given by:

$$P\,(z_{d_n}^{h} = \mathrm{k}|z_{..}^{(m)}, z_{-dn}^{(h)}, w_{..}^{(m)}, w_{..}^{(h)}, t_{..}^{(.)}\,) \propto \frac{\beta_h + c_{-dn,k}^{w_{dn}^{(h)}}}{vh\beta h + c_{-dn,k}^{(h)}} \; \frac{\alpha + c_{-dn,k}^{(d_b)} + c_{.,k}^{(dm)}}{K\alpha + N_{dm} + N_{db} - 1}$$

$$\times \; \frac{t_d^{\psi_{k1}-1}(1-t_d)^{\psi_{k2}-1}}{B(\psi_{k1},\psi_{k2})} \qquad (1)$$

Type equation here.

(h) where denotes the number of hashtags type $w_{dn}^{(b)}$ assigned to latent topic k, excluding the hashtag currently undergoing processing; $c_{-dn,k}^{(h)}$ denotes the number of hashtags assigned to latent topic k, excluding the assignment at position d_n*;* $c_{-dn,k}^{(dh\ \ )}$ denotes the number of hashtags assigned to latent topic k in tweet d, excluding the hashtag currently undergoing processing; $c_{.,k}^{(dm)}$ denotes the number of words assigned to latent topic k in tweet *d;* V_b is the number of unique hashtags; t_d is the tweet time stamp omitting position subscripts and superscripts (all words and hashtags share the same time stamp); ψ_{k1}, ψ_{k2} are the parameters of the beta distribution for latent topic k.

The probability for a hashtag given the observed words and time stamps is:

$$p\left(w_{vn}^{h} \middle| w_{v.}^{(m)}, \mathrm{t_v}\right) = \int p\left(w_{vn}^{(h)} \middle| \theta^{(v)}\right) p\left(\theta^{(v)} \middle| w_{v.}^{|(m)}, t_v\right) d\theta^{(v)}$$

$$s \approx \frac{1}{\acute{S}} \sum_{s=1}^{\acute{S}} \quad \sum_{k=1}^{K} \emptyset_{h,k,w_{vn}^{(h)}}^{(s)} \theta_k^{(v)(s)}, \qquad (2)$$

where $\acute{S}$ is the total number of recorded sweeps, and the superscript ***s*** marks the parameters computed based on a specific recorded sweep. To provide the top ***N*** predictions, they ranked $p\left(w_{vn}^{(h)} \middle| .\right)$ from largest to smallest and output the first ***N*** hashtags.

rLDA , the authors introduced a novel probabilistic formulation to obtain the relevance of a tag with considering all the other images and their tags and also they proposed a novel model called regularized latent Dirichlet allocation (rLDA). This model can estimates the latent topics for each document, with making use of other documents. They used a collective inference scheme to estimate the distribution of latent topics and applied a deep network structure to analyze the benefit of regularized LDA (Wang et al., 2014a) (Lu and Lee, 2015).

Table1 4. Some impressive articles based on LDA: between 2014-2015

Author-study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Rao et al., 2014)	emotion LDA (ELDA	2014	Gibbs sampling	LDA	Social emotion classification of online news
(Kim and Shim, 2014)	TWILITE	2014	EM algorithm	LDA	Recommendation system for Twitter
(Cohen et al., 2014)	Red-LDA	2014	Gibbs-Samplin	-LDA	Extract information and and data modeling… in Patient Record Notes
(Cohen et al., 2014)	Biterm Topic Modeling (BTM)	2014	Gibbs sampling	LDA	Document clustering for short text
(Yang and Rim, 2014)	Trend Sensitive-Latent Dirichlet Allocation (TS-LDA)	2014	Gibbs sampling	LDA - normalized Discounted Cumulative Gain (nDCG) - Amazon Mechanical Turk (AMT)2 platform	Interesting tweets discover for users, system recommendation
(Wang et al., 2014c)	Fine-grained Labeled LDA (FL-LDA), Unified Fine-grained Labeled LDA (UFL-LDA)	2014	Gibbs sampling	LDA	Aspect extraction and review mining
(Wang et al., 2014a) (Lu and Lee, 2015)	regularized latent Dirichlet allocation (rLDA)	2014	Variational Bayes inference	LDA	Automatic image tagging or tag recommendation
(Cheng et al., 2014)	generative probabilistic aspect mining model (PAMM)	2014	Expectation-maximization (EM)	LDA	Opinion mining and groupings of drug reviews, aspect mining
(Zheng et al., 2014)	AEP-based Latent Dirichlet Allocation (AEP-LDA)	2014	Gibbs sampling	LDA	Opinion /aspect mining and sentiment word identification
(Bagheri et al., 2014)	ADM-LDA	2014	Gibbs sampling	LDA Markov chain	Aspect mining and sentiment analysis
(Xie et al., 2015)	MRF-LDA	2015	EM algorithm	-LDA -Markov Random Field	Exploiting word correlation knowledge
	Hawkes-LDA	2015	Variational Bayes inference	LDA	Analyzing text content and modeling scientific influence
(Yuan et al.,	LightLDA	2015	Gibbs	LDA	Topic modeling

2015)			sampling	Metropolis-Hastings sampling algorithm	for very large data sizes
(Nguyen et al., 2015)	Latent Feature LDA (LF-LDA), LF-DMM	2015	Gibbs sampling	LDA	Document clustering for short text
	TH Rank	2015	Gibbs sampling	-LDA -Author-Conference-Topic (ACT) model - PageRank	Topic sensitive ranking and find the relevant papers in journals
(Li et al., 2015a)	author-topic-community (ATC)	2015	Expectation-maximization (EM)	LDA	Author community discovery and Author interest profiling
(Lu and Lee, 2015)	TOT-MMM	2015	Gibbs sampling	LDA	Twitter Hashtag Recommendation
(Yu et al., 2015b)	link-field-topic (LFT),	2015	Gibbs sampling	LDA - semantic link weight (SLW)	Semantic community detection and dynamic topic discovery
(Jiang et al., 2015)	Scalable Geographic Web Search Topic Discovery (SG-WSTD)	2015	Gibbs sampling	LDA k-means algorithm	Geographic web search topic discovery and extracting geographic information
(Fu et al., 2015)	dynamic NJST (dNJST)	2015	Gibbs sampling algorithms	-LDA -hierarchical Dirichlet process (HDP)	Dynamic sentiment topic discovery in Chinese social media
(Li et al., 2015b)	Frequency-LDA (FLDA) and Dependency-Frequency-LDA (DFLDA)	2015	Gibbs sampling algorithms	-LDA	Multi-label document categorization

2.2.4. A brief look from some impressive past works: Research in 2016

According to Table 5, some of the popular works published for this year focused on a variety of topics, such as recommendation system(Cheng and Shen, 2016) (Zoghbi et al., 2016) (Zhao et al., 2016) , opinion mining and aspect mining (Bagheri et al., 2014) (Zheng et al., 2014) (Cheng et al., 2014) (Wang et al., 2014c).

A bursty topic on Twitter is one that triggers a surge of relevant tweets within a short period of time, which often reflects important events of mass interest. How to leverage Twitter for early detection of bursty topics has, therefore, become an important research problem with immense practical value. In **TopicSketch**(Xie et al., 2016), proposed a sketch-based topic model together with a set of techniques to achieve real-time bursty topic detection from the perspective of topic modeling, that called in this paper TopicSketch.

The bursty topics are often triggered by some events such as some breaking news or a compelling basketball game, which get a lot of attention from people, and "**force**" people to tweet about them intensely. For example, in physics, this "**force**" can be expressed by "**acceleration**", which in our setting describes the change of "**velocity**", i.e., arriving rate of tweets. Bursty topics can get significant acceleration when they are bursting, while the general topics usually get nearly zero acceleration. So the "***acceleration***" trick can be used to preserve the information of bursty topics but filter out the others. Equation (3) shows how we calculate the "**velocity**" $\hat{v}(t)$ and "**acceleration**" $\hat{\alpha}(t)$ of words.

$$\hat{v}_{\Delta T} = \sum_{t_i \leq t} X_i \, . \frac{\exp((t_i - t)/\Delta T)}{\Delta T}, \; \hat{\alpha}(t) = \frac{\hat{v}_{\Delta T2}(t) - \hat{v}_{\Delta T1}(t)}{\Delta T_1 - \Delta T_2}. \qquad (3)$$

In Equation (1), X_i is the frequency of a word (or a pair of words, or a triple of words) in the *i-th* tweet, t_i is its timestamp. The exponential part in $\hat{v}_{\Delta T}(t)$ works like a soft moving window, which gives the recent terms high weight, but gives low weight to the ones far away, and the smoothing parameter ΔT is the window size.

Hashtag-LDA, the authors a personalized hashtag recommendation approach is introduced according to the latent topical information in untagged microblogs. This model can enhance the influence of hashtags on latent topics' generation by jointly modeling hashtags and words in microblogs. This approach inferred by Gibbs sampling to latent topics and considered a real-world Twitter dataset to evaluation their approach (Zhao et al., 2016). **CDLDA** proposed a conceptual dynamic latent Dirichlet allocation model for tracking and topic detection for conversational communication, particularly for spoken interactions. This model can extract the dependencies between topics and speech acts. The CDLDA applied hypernym information and speech acts for topic detection and tracking in conversations, and it captures contextual information from transitions, incorporated concept features and speech acts (Yeh et al., 2016).

Tablel 5. Some impressive articles based on LDA for 2016

Author-Study	Model	Years	Parameter Estimation / Inference	Methods	Problem Domain
(Zhao et al., 2016)	Hashtag-LDA	2016	Gibbs sampling	LDA	Hashtag recommendation

					and Find relationships between topics and hashtags
(Hong et al., 2016)	PMB-LDA	2016	Expectation-maximization (EM)	LDA	Extract the population mobility behaviors for large scale
(Lee et al., 2016)	Automatic Rule Generation (LARGen	2016	Gibbs Sampling	LDA	Malware analysis and Automatic Signature Generation
(Liu et al., 2016)	PT-LDA	2016	Gibbs-EM algorithm	LDA	Personality recognition in social network
(Li et al., 2016a)	Corr-wddCRF	2016	Gibbs sampling	LDA	Knowledge Discovery in Electronic Medical Record
(Zoghbi et al., 2016)	multi-idiomatic LDA model (MiLDA)	2016	Gibbs sampling	LDA bilingual LDA (BiLDA	Content-based recommendation and automatic linking
(Cheng and Shen, 2016)	Location-aware Topic Model (LTM)	2016	Gibbs sampling	LDA	Music Recommendation
(Miao et al., 2016)	TopPRF	2016	Gibbs sampling	LDA	Evaluate the relevancy between feedback documents
(Rao, 2016)	contextual sentiment topic model (CSTM)	2016	Expectation-maximization (EM)	LDA	Emotion classification in social network
(Yeh et al., 2016)	conceptual dynamic latent Dirichlet allocation (CDLDA)	2016	Gibbs sampling	LDA	Topic detection in conversations
(Lu et al., 2016)	multiple-channel latent Dirichlet allocation (MCLDA)	2016	Gibbs sampling	LDA	Find the relations between diagnoses and medications from healthcare data
(Qian et al., 2016)	multi-modal event topic model (mmETM)	2016	Gibbs sampling	LDA	Tracking and social event analysis
(Fu et al., 2016)	Dynamic Online Hierarchical Dirichlet Process model (DOHDP)	2016	Gibbs samplin	LDA	Dynamic topic evolutionary discovery for Chinese social media
(Xie et al., 2016)	Topicsketch	2016	Gibbs sampling	LDA - tensor decomposition algorithm - Count-Min algorithm	Realtime detection and bursty topics dicovery from Twitter

(Zeng et al., 2016)	fast online EM (FOEM)	2016	Expectation-maximization (Batch EM)	LDA	Big topic modeling
(Alam et al., 2016)	Joint Multi-grain Topic Sentiment (JMTS)	2016	Gibbs sampling	LDA	Extracting semantic aspects from online reviews
(Qin et al., 2016)	character–word topic model (CWTM)	2016	Gibbs sampling	LDA	Capture the semantic contents in text documents(Chinese language).

2.2 Topic Modeling for which the area is used?

With the passage of time, the importance of Topic modeling in different disciplines will be increase. According to previous studies, we present a taxonomy of current approaches topic models based on LDA model and in different subject such as Social Network(McCallum et al., 2005) (Wang et al., 2013) (Henderson and Eliassi-Rad, 2009, Yu et al., 2015a), Software Engineering(Linstead et al., 2008) (Chen et al., 2012) (Gethers and Poshyvanyk, 2010) (Linstead et al., 2007), Crime Science(Chen et al., 2015) (Gerber, 2014) (Wang et al., 2012) and also in areas of Geographical(Cristani et al., 2008) (Yin et al., 2011) (Sizov, 2010) (Tang et al., 2013), Political Science(Greene and Cross, 2015) (Cohen and Ruths, 2013) , Medical/Biomedical (Liu et al., 2010) (Huang et al., 2013) (Wang et al., 2011) (Zhang et al., 2017) (Xiao et al., 2017) and Linguistic science (Bauer et al., 2012) (McFarland et al., 2013) (Eidelman et al., 2012) (Wilson and Chew, 2010) (Vulić et al., 2011) as illustrated by Fig. 2.

Fig2. A clear vision of the application of Topic modeling in various sciences (Based on previous work**).**

A. Topic modeling in Linguistic science

LDA is an advanced textual analysis technique grounded in computational linguistics research that calculates the statistical correlations among words in a large set of documents to identify and quantify the underlying topics in these documents. In this subsection, we examine some of topic modeling methodology from computational linguistic research. **Vulic et al.** employed the distributional hypothesis in various direction and it efforts to cancel the requirement of a seed lexicon as an essential prerequisite for use of bilingual vocabulary and introduce various ways to identify the translation of words among languages (Vulić et al., 2011). **Eidelman et al.** introduced a method that leads the machine translation systems to relevant translations based on topic-specific contexts and used the topic distributions to obtain topic-dependent lexical weighting probabilities. They considered a topic model for training data, and adapt the translation model. To evaluate their approach, they performed experiments on Chinese to English machine translation and show the approach can be an effective strategy for dynamically biasing a statistical machine translation towards relevant translations (Eidelman et al., 2012).

Table6. Impressive works LDA-based in Linguistic science

Study- Author	Year	Purpose	Dataset
(Vulić et al., 2011)	2011	Introduce various ways to identify the translation of words among languages [BiLDA].	A Wikipedia dataset (Arabic, Spanish, French, Russian and English)
(Wilson and Chew, 2010)	2010	Obtain term weighting based on LDA	A multilingual dataset
(McFarland et al., 2013)	2013	Present a diversity of new visualization techniques to make concept of topic-solutions	Dissertation abstracts_1980–2010 - 1 million abstracts
(Bauer et al., 2012)	2012	A topic modeling approach, that it consider geographic information	Foursquare Dataset
(Lui et al., 2014)	2014	An approach that is capable to find a document with different language	ALTW2010
(Heintz et al., 2013)	2013	A method for linguistic discovery and conceptual metaphors resources	Wikipedia

McFarland et al. presented a diversity of new visualization techniques to make the concept of topic-solutions and introduce new forms of supervised LDA, to evaluation they considered a corpus of dissertation abstracts from 1980–2010 that belongs to 240 universities in the United States(McFarland et al., 2013). **Bauer et al.** developed a standard topic modeling approach, that it consider geographic and temporal information and this approach used to Foursquare data and discover the dominant topics in the proximity of a city. Also, the researchers have shown that the abundance of data available in location-based social network (LBSN) enables such models to obtain the topical dynamics in urbanite environments(Bauer et al., 2012). **Heintz et al** have introduced a method for discovery of linguistic and conceptual metaphors resources and built an LDA model on Wikipedia; align its topics to possibly source and aim concepts, they used from both target and source domains to identify sentences as potentially metaphorical(Heintz et al., 2013). **Lui et al.** presented an approach that is capable to find a document with a different language and identify the current language in a document and next step calculate their relative proportions, this approach is based on LDA and used from ALTW2010 as a dataset to evaluation their method (Lui et al., 2014).

B. Topic modeling in political science

Some topic modeling methods have been adopted in the political science literature to analyze political attention. In settings where politicians have limited time-resources to express their views, such as the plenary sessions in parliaments, politicians must decide what topics to address. Analyzing such speeches can thus provide insight into the political priorities of the politician under consideration. Single membership topic models that assume each speech relates to one topic; have successfully been applied to plenary speeches made in the 105th to the 108th U.S. Senate in order to trace political attention of the Senators within this context over time [18]. Also, some researchers proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time and identify both niche topics related to events at a particular point in time and broad, long-running topics. This paper has focused on European Parliament speeches, the proposed topic modeling method has a number of potential applications in the study of politics, including the analysis of speeches in other parliaments, political manifestos, and other more traditional forms of political texts (Greene and Cross, 2015). **Cohen et al.** The purpose of the study is to examine the various effects of dataset selection with consideration of policy orientation classifiers and built three datasets that each data set include of a collection of Twitter users who have a political orientation. In this approach, the output of an LDA has been used as one of many features as a fed for apply SVM classifier and another part of this method used an LLDA that Considered as a stand-alone classifier. Their assessment showed that there are some limitations to building labels for non-political user categories (Cohen and Ruths, 2013).

Table7. Impressive works LDA-based in political science

Study- Author	**Year**	**Purpose**	**Dataset**
(Cohen and Ruths, 2013)	2013	Evaluate the behavioral effects of different databases from political orientation classifiers	-Political Figures Dataset -Politically Active Dataset (PAD) -Politically Modest Dataset (PMD) -Conover 2011 Dataset (C2D)
(Fang et al., 2012)		Introduce a topic model for contrastive opinion modeling	-Statement records of U.S. senators
(Balasubramanyan et al., 2012)	2012	Detection topics that evoke different reactions from communities that lie on the political spectrum	A collection of blog posts from five blogs: 1. Carpetbagger(CB)- thecarpetbaggerreport.com 2. Daily Kos(DK) - dailykos.com 3. Matthew Yglesias(MY) - yglesias.thinkprogress.org

			4.Red State(RS) - redstate.com 5.Right Wing News(RWN) - rightwingnews.com
(Chen et al., 2010)	2010	Discover the hidden relationships between opinion word and topics words	The statement records of senators through the Project Vote Smart (http://www.votesmart.org)
(Song et al., 2014)	2014	Analyze issues related to Korea's presidential election	Project Vote Smart WebSite (https://votesmart.org/)
(Levy and Franklin, 2014)	2014	Examine Political Contention in the U.S. Trucking Industry	Regulations.gov online portal
(Zirn and Stuckenschmidt , 2014)	2015	presented a method for multi-dimensional analysis of political documents	three Germannational elections (2002, 2005 and 2009)

Fang et al. They suggested a new unsupervised topic model based on LDA for contrastive opinion modeling which purpose to find the opinions from multiple views, according to a given topic and their difference on the topic with qualifying criteria, the model called Cross-Perspective Topic (CPT) model. They performed experiments with both qualitative and quantitative measures on two datasets in the political area that include: first dataset is statement records of U.S. senators that show political stances of senators by these records, also for the second dataset, extracted of world News Medias from three representative media in U.S (New York Times), China (Xinhua News) and India (Hindu). To evaluate their approach with other models used corrIDA and LDA as two baselines (Fang et al., 2012). **Yano et al.** applied several probabilistic models based on LDA to predict responses from political blog posts.in more detail, they used topic models LinkLDA and CommentLDA to generate blog data(topics, words of post) in their method and with this model can found a relationship between the post, the commentators and their responses. To evaluation, their model gathered comments and blog posts with focusing on American politics from 40 blog sites (Yano et al., 2009, Yano and Smith, 2010).

Madan et al. Introduced a new application of universal sensing based on using mobile phone sensors and used an LDA topic model to discover pattern and analysis of behaviors of people who changed their political opinions, also evaluated to various political opinions for residents of individual , with consider a measure of dynamic homophily that reveals patterns for external political events. To collect data and apply their approach, they provided a mobile sensing platform to capture social interactions and dependent variables of American Presidential campaigns of John McCain and President Barack Obama in last three months of

2008 (Madan et al., 2011). **Balasubramanyan et al.** they analyzed reactions of emotional and suggested a novel model Multi Community Response LDA (MCR-LDA) which in fact is a multi-target and for predicting comment polarity from post content used sLDA and support vector machine classification (Balasubramanyan et al., 2012). To evaluation, their approach they provided a dataset of blog posts from five blogs that focus on US politics that was made by (Yano et al., 2009).

Chen et al. suggested a generative model to auto discover of the latent associations between opinion words and topics that can be useful for extraction of political standpoints and used an LDA model to reduce the size of adjective words, the authors successfully get that sentences extracted by their model and they shown this model can effectively in different opinions. They were focused on statement records of senators that includes 15, 512 statements from 88 senators from Project Vote Smart WebSite (Chen et al., 2010). **Song and et al.** It was examined how social and political issues related to South Korean presidential elections in 2012 on Twitter and used an LDA method to evaluate the relationship between topics extracted from events and tweets (Song et al., 2014). **Zirn et al.** proposed a method for evaluating and comparing documents, based on an extension of LDA, and used LogicLDA and Labeled LDA approaches for topic modeling in their method. They are considered German National Elections since 1990 as a dataset to apply their method and shown that the use of their method consistently better than a baseline method that simulates manual annotation based on text and keywords evaluation (Zirn and Stuckenschmidt, 2014).

C. Topic modeling in Medical/Biomedical

Topic models applied to text mining in Medical/biomedical domain, according to previous studies, LDA can be very effective and functional in this field. Topic modeling could be advantageously applied to the large datasets of biomedical/medical research. For example, a group of researchers, introduced three LDA-like models and found that this model cans higher accuracy than the state-of-the-art alternatives. Authors demonstrated that this approach based on LDA could successfully recognize the probabilistic patterns between Adverse drug reaction (ADR) topics and used ADRS database for evaluation their approach. The aim of the authors to predict ADR from a large number of ADR candidates to obtain a drug target(Xiao et al., 2017). **Zhang et al.** They focused on the issue of professionalized medical recommendations and proposed a new healthcare recommendation system that called iDoctor, that used Hybrid matrix factorization methods for the professionalized doctor recommendation. In fact, They adopted an LDA topic model to extract the topics of doctor features and analyzing document similarity. The dataset this article is college from a crowd sourced website that called Yelp. Their result showed that iDoctor can increase the accuracy of health recommendations and it can has higher prediction in users ratings(Zhang et al., 2017).

Table8. Impressive works LDA-based in medical/biomedical

Study-Author	**Year**	**Purpose/problem domain**	**Dataset**

(Xiao et al., 2017)	2017	Presented three LDA-based models Adverse Drug Reaction Prediction	ADReCS database
(Wang et al., 2011)	2011	Extract biological terminology	-MEDLINE and Bio-Terms Extraction -Chem2Bio2Rdf.
(Zhang et al., 2017)	2017	User preference distribution discovery and identity distribution of doctor feature	-Yelp Dataset (Yelp.com)
(Wu et al., 2012b)	2012	-Ranking GENE-DRUG -Detecting relationship between gene and drug	National Library of Medicine
(Paul and Dredze, 2011)	2011	Analyzing public health information on Twetter	20 disease articles of twitter data
(Wang et al., 2013)	2013	Analysis of Generated Content by User from social networking sites	one million English posted from Facebook's server logs
(Huang et al., 2013)	2013	Pattern discovery and extraction for Clinical Processes	a data-set from Zhejiang Huzhou Central Hospital of China
(Liu et al., 2010)	2011	Identifying miRNA-mRNA in functional miRNA regulatory modules	mouse mammary dataset (Zhu et al., 2010)
(Zhang et al., 2011)	2011	Extract common relationship	T2DM Clinical Dataset
	2012	Extract the latent topic in	T2DM Clinical Dataset

(Jiang et al., 2012)		Traditional Chinese Medicine document	

Wang et al. they suggested an approach based on LDA that called Bio-LDA that can identify biological terminology to obtain latent topics. The authors have shown that this approach can be applied in different studies such as association search, association predication, and connectivity map generation. And they showed that Bio-LDA can be applied to increase the application of molecular bonding techniques as heat maps(Wang et al., 2011). **Wu et al.** proposed a topic modeling for rating gene-drug relations by using probabilistic KL distance and LDA that called LDA-PKL and showed that the suggested model achieved better than Mean Average Precision (MAP). They found that the presented method achieved a high Mean Average Precision (MAP) to rating and detecting pharmacogenomics(PGx) relations. To analyze and apply their approach used a dataset from National Library of Medicine(Wu et al., 2012b). **Paul et al.** Presented Ailment Topic Aspect Model (ATAM) to the analysis of more than one and a half million tweets in public health and they were focused on a specific question and specific models; “what public health information can be learned from Twitter?(Paul and Dredze, 2011)”.

Huang et al. they introduced an LDA based method to discover patterns of internal treatment for Clinical processes (CPs), and currently, detect these hidden patterns is one of the most serious elements of clinical process evaluation. Their main approach is to obtain care flow logs and also estimate hidden patterns for the gathered logs based on LDA. Patterns identified can apply for classification and discover clinical activities with the same medical treatment. To experiment the potentials of their approach, used a data-set that collected from Zhejiang Huzhou Central Hospital of China(Huang et al., 2013). **Liu et al.** They introduced a model for the discovery of functional miRNA regulatory modules (FMRMs) that merge heterogeneous datasets and it including expression profiles of both miRNAs and mRNAs, using or even without using exploit the previous goal binding information. This model used a topic model based on Correspondence Latent Dirichlet Allocation (Corr-LDA). As an evaluation dataset, they perform their method to mouse model expression datasets to study the issue of human breast cancer. The authors found that their model is mighty to obtain different biologically meaningful models (Liu et al., 2010). **Zhang et al.** The authors had a study on Chinese medical (CM) diagnosis by topic modeling and introduced a model based on Author-Topic model to detect CM diagnosis from Clinical Information of Diabetes Patients, and called Symptom-Herb-Diagnosis topic (SHDT) model. Evaluation dataset has been collected from 328 diabetes patients. The results indicated that the SHDT model can discover herb prescription topics and typical symptom for a bunch of important medical-related diseases in comorbidity diseases (such as; heart disease and diabetic kidney)(Zhang et al., 2011).

D. Topic modeling in Geographical/locations

There is a significant body of research on geographical topic modeling. According to past work, researchers have shown that topic modeling based on location information and textual information can be effective to discover geographical topics and Geographical Topic Analysis. **Yin et al.** This article examines the issue of topic modeling to extract the topics from geographic information and GPS-related documents. They suggested a new location text method that is a combination of topic modeling and geographical clustering called LGTA

(Latent Geographical Topic Analysis). To test their approaches, they collected a set of data from the website Flickr, according to various topics(Yin et al., 2011). **Sizov et al.** They introduced a novel method based on multi-modal Bayesian models to describe social media by merging text features and spatial knowledge that called GeoFolk. As a general outlook, this method can be considered as an extension of Latent Dirichlet Allocation (LDA). They used the available standard CoPhIR dataset that it contains an abundance of over 54 million Flickr. The GeoFolk model has the ability to be used in quality-oriented applications and can be merged with some models from Web 2.0 social (Sizov, 2010)**.** **Tang et al.** they proposed a multiscale LDA model that is a combination of multiscale image representation and probabilistic topic model to obtain effective clustering VHR satellite images (Tang et al., 2013).

Table 9. Impressive works LDA-based in geographical/locations

Study-Author	**Year**	**Purpose**	**Dataset**
(Sizov, 2010)	2010	Discovering multi-faceted summaries of documents	CoPhIR dataset
(Yin et al., 2011)	2011	Content management and retrieval	Flicker Dataset
(Tang et al., 2013)	2013	Semantic clustering in very high resolution panchromatic satellite images	A QUICKBIRD image of a suburban area
(Eisenstein et al., 2010)	2010	Data Discovery, Evaluation of geographically coherent linguistic regions and find the relationship between topic variation and regional.	A Twitter Dataset
(Cristani et al., 2008)	2008	Geo-located image categorization and georecognition	3013 images Panoramio in France
(Zhang et al., 2015)	2015	Cluster discovery in geo-locations	Reuters-21578
(McInerney and Blei, 2014)	2014	Discovering newsworthy information From Twitter	A small Twitter Dataset

Eisenstein et al. They introduced a model that includes two sources of lexical variation: geographical area and topic, in another word, this model can discover words with geographical coherence in different linguistic regions, and find a relationship between regional and variety of topics. To test their model, they gathered a dataset from the website Twitter and also we can say that, also can show from an author's geographic location from raw text (Eisenstein et al., 2010) (Tang et al., 2013) (Yin et al., 2011) (Sizov, 2010). **Cristani et al.** They suggested a statistical model for classification of geo-located images based on latent representation. In this model, the content of a geo-located database able be visualized by means of some few selected images for each geo-category. This model can be considered as an extension of probabilistic Latent Semantic Analysis (pLSA). They built a database of the geo-located image which contains 3013 images (Panoramio), that is related to southeastern France (Cristani et al., 2008).

Zhang et al. In this work, Authors focused on the issue of identifying textual topics of clusters including spatial objects with descriptions of text. They presented combined methods based on cluster method and topic model to discover textual object clusters from documents with geo-locations. In fact, they used a probabilistic generative model (LDA) and the DBSCAN algorithm to find topics from documents. In this paper, they utilized dataset Reuters-21578 as a dataset for Analysis of their methods (Zhang et al., 2015). **McInerney et al.** they presented a study on characterizing significant reports from Twitter, The authors, introduced a probabilistic model to topic discovery in the geographical topic area and this model can find hidden significant events on Twitter and also considered stochastic variational inference (SVI) to apply gradient ascent on the variable objective with LDA. They collected 2,535 geo-tagged tweets from the Upper Manhattan area of New York. that the KL divergence is a good metric to identifying the significant tweet event, but for a large dataset of news articles, the result will be negative(McInerney and Blei, 2014).

E. Software engineering and topic modeling

Software evolution and source code analysis can be effective in solving current and future software engineering problems. Topic modeling has been used in information retrieval and text mining where it has been applied to the problem of briefing large text corpora. Recently, many articles have been published for evaluating / mining software using topic modeling based on LDA. **Linstead et al.** For the first time, they used LDA, to extract topics in source code and perform to visualization of software similarity, In other words, LDA uses an intuitive approach for calculation of similarity between source files with obtain their respective distributions of each document over topics. They utilized their method on 1,555 software projects from Apache and SourceForge that includes 19 million source lines of code (SLOC). The authors demonstrated this approach, can be effective for project organization, software refactoring (Linstead et al., 2007). **Gethers et al.** They introduced a new coupling metric based on Relational Topic Models (RTM) that called Relational Topic based Coupling (RTC), that can identifying latent topics and analyze the relationships between latent topic distributions software data. Also, can say that the RTM is an extension of LDA. The authors used thirteen open source software systems for evaluation this metric and demonstrated that RTC has a useful and valuable impact on the analysis of large software systems(Gethers and Poshyvanyk, 2010).

Asuncion et al. the authors focused on software traceability by topic modeling and proposed a combining approach based on LDA model and automated link capture. They utilized their method to several data sets and demonstrated how topic modeling increase software traceability, and found this approach, able to scale for carried larger numbers from artifacts (Asuncion et al., 2010). **Thomas et al.** They studied about the challenges use of topic models to mine software repositories and detect the evolution of topics in the source code, and suggested the apply of statistical topic models (LDA) for the discovery of textual repositories. Statistical topic models can have different applications in software engineering such as bug prediction, traceability link recovery and software evolution (Thomas, 2011). **Chen et al.** used a generative statistical model(LDA model) for analyzing source code evolution and find relationships between software defects and software development. They showed LDA can easily scale to large documents and utilized their approach on three large dataset that includes: Mozilla Firefox, and Mylyn, Eclipse (Chen et al., 2012). **Linsteadet al.** used and utilized Author-Topic models(AT) to analysis in source codes. AT modeling is an extension of LDA model that evaluation and obtain the relationship of authors to topics and applied their method on Eclipse 3.0 source code including of 700,000 code lines and 2,119 source files with considering of 59 developers. They demonstrated that topic models provided the effective and statistical basis for evaluation of developer similarity(Linstead et al., 2008).

Tian et al. introduced a method based on LDA for automatically categorizing software systems, called LACT. For evaluation of LACT, used 43 open-source software systems in different programming languages and showed LACT can categorization of software systems based on the type of programming language (Tian et al., 2009). **Lukinet al.** Proposed an approach topic modeling based on LDA model for the purpose of bug localization. Their idea, applied to the analysis of same bugs in Mozilla and Eclipse and result showed that their LDA-based approach is better than LSI for evaluate and analyze of bugs in these source codes (Lukins et al., 2008, Lukins et al., 2010). **Yang et al.** They introduced a topic-specific approach by considering the combination of description and sensitive data flow information and used an advanced topic model based on LDA with GA, to understanding malicious apps, cluster apps according to their descriptions. They utilized their approach on 3691 benign and 1612 malicious application. The authors found Topic-specific, data flow signatures are very efficient and useful in highlighting the malicious behavior (Yang et al., 2017).

Table10. Impressive works LDA-based in software engineering

Study- Author	Year	Purpose	Dataset
(Linstead et al., 2007)	2007	Mining software and extracted concepts from code	SourceForge and Apache(d 1,555 projects)
(Gethers and Poshyvanyk, 2010)	2010	Identifying latent topics and find their relationships in source code	Thirteen open source software systems
(Asuncion et al., 2010)	2010	Generating traceability links	ArchStudio software project
(Chen et al., 2012)	2012	Find relationship between the conceptual concerns in source code.	source code entities
(Linstead et al., 2008)	2008	Analyzing Software Evolution	, open source Java projects, Eclipse and ArgoUML

(Tian et al., 2009)	2009	Automatic Categorization of Software systems	43 open-source software systems
(Lukins et al., 2008)	2008	Source code retrieval for bug localization	Mozila, Eclipse source code
(Lukins et al., 2010)	2010	Automatic bug localization and evaluate its effectiveness	Open source software such as (Rhino, and Eclipse)
(Yang et al., 2017)	2017	Detection of malicious Android apps	1612 malicious application

F. Topic modeling in Social Network / Microblogs (such as Twitter)

Social networks are a rich source for knowledge discovery and behavior analysis. For example, Twitter is one of the most popular social networks that its evaluation and analysis can be very effective for analyzing user behavior and etc. Recently, researchers have proposed many LDA approaches to analyzing user tweets on Twitter. **Weng et al.** In this paper, the authors were concentrated on identifying influential twitterers on Twitter and proposed an approach based on an extension of PageRank algorithm to rate users, called TwitterRank, and used an LDA model to find latent topic information from a large collection of documentation. For evaluation this approach, they prepared a dataset from Top-1000 Singapore-based twitterers, showed that their approach is better than other related algorithms (Weng et al., 2010). **Hong et al.** This paper examines the issue of identifying the Message popularity as measured based on the count of future retweets and sheds. The authors utilized TF-IDF scores and considered it as a baseline, also used Latent Dirichlet Allocation (LDA) to calculate the topic distributions for messages. They collected a dataset that includes 2,541,178 users and 10,612,601 messages and demonstrated that this method can identify messages which will attract thousands of retweets (Hong et al., 2011).

Bhattacharya et al. In this paper, they focused on topical recommendations on tweeter and presented a novel methodology for topic discovery of interests of a user on Twitter. In fact, they used a Labeled Latent Dirichlet Allocation (L-LDA) model to discover latent topics between two tweet-sets. The authors found that their method could be better than content based methods for discovery of user-interest (Bhattacharya et al., 2014). **Kim et al.** They suggested, a recommendation system based on LDA for obtaining the behaviors of users on Twitter, called TWILITE. In more detail, TWILTW can calculate the topic distributions of users to tweet messages and also they introduced ranking algorithms in order to recommend top-K followers for users on Twitter (Kim and Shim, 2014). **Wang et al.** They investigated in the context of a criminal incident prediction on Twitter. They suggested an approach for analysis and understanding of Twitter posts based a probabilistic language model and also considered a generalized linear regression model. Their evaluation showed that this approach is the capability of predict hit-and-run crimes, only using information that exists in the content of the training set of tweets (Wang et al., 2012).

Tablel11. Impressive works LDA-based in social network

Study- Author	Year	Purpose	Dataset
Weng et all (Weng et al., 2010)	2010	Finding influential twitterers on social network(Twitter)	Top-1000 Singapore-based twitterers
Bhattacharya et all (Bhattacharya et al., 2014)	2014	Building a topical recommendation systems	A twitter dataset
Kim et all (Kim and Shim, 2014)	2014	A recommendation system for Twitter	A twitter dataset
Cordeiro et all(Cordeiro, 2012)	2012	Analysis and discovered events on Twitter	A twitter dataset
Tan et all (Tan et al., 2014)	2014	Analyze public sentiment variations regarding a certain tar on Twitter	A twitter dataset
Roberts et all (Roberts et al., 2012)	2012	Analysis of the emotional and stylistic distributions on Twitter	A twitter dataset
Ren et all (Ren et al., 2016)	2016	A topic-enhanced word embedding for Twitter sentiment classification	SemEval-2014
Li et al (Li et al., 2016b)	2016	Categorize emotion tendency on Sina Wibo	A Sina Wibo dataset

Godin et al. In this paper, they introduced a novel method based LDA model to hashtag recommendation on Twitter that can categories posts with them (hashtags)(Godin et al., 2013). **Lin et al.** They investigated the cold-start issue with useing the social information for App recommendation on Twitter and used an LDA model to discovering latent group from "Twitter personalities" to recommendations discovery. For test and experiment, they considered Apple's iTunes App Store and Twitter as a dataset. Experimental results show, their approach significantly better than other state-of-the-art recommendation techniques (Lin et al., 2013).

Cordeiro et al. presented a technique to analysis and discovered events by an LDA model. Authors found that this method can detect events in inferred topics from tweets by wavelet analysis. For test and evaluation, they collected 13.6 million tweets from Twitter as a dataset and showed the use of both hashtag names and inferred topics is a beneficial effect in description information for events (Cordeiro, 2012). **Pier et al.** In this paper, they investigated the issue of how to effectively discovery and find health-related topics on Twitter and presented an LDA model for identifies latent topic information from a dataset and it includes 2,231,712 messages from 155,508 users. They found that this method may be a valuable tool for detect public health on Twitter (Prier et al., 2011). **Tan and et al.** focused on tracking public sentiment and modeling on Twitter. They suggest a topic model approach based on LDA, Foreground and Background LDA to distill topics of the foreground. Also proposed another method for can rank a set of reason candidates in natural language, called Reason Candidate and Background LDA (RCB-LDA). Their results showed that their models can be used to identify special topics and find different aspects (Tan et al., 2014). **Roberts et al.** collected a large corpus from Twitter in seven emotions that includes; disgust, Anger, Fear, Love, Joy, sadness, and surprise. They used a probabilistic topic model, based on LDA, which considered for discovery of emotions in a corpus of Twitter conversations(Roberts et al., 2012). **Srijith et al.** This paper proposed a probabilistic topic model based on hierarchical Dirichlet processes (HDP)) for detection of sub-story. They compared HDP with spectral clustering (SC) and locality sensitive hashing (LSH) and showed that HDP is very effective for story detection data sets, and has an improvement of up to 60% in the F-score (Srijith et al., 2017).

Ren et al. proposed a method based on Twitter sentiment classification using topic-enhanced word embedding and also used an LDA model to generate a topic distribution of tweets, considered SVM for classifying tasks in sentiment classification. They used the dataset on SemEval-2014 from Twitter Sentiment Analysis Track. Experiments show that their model can obtain 81.02% in macro F-measure (Ren et al., 2016). **Wang et al.** focused on examining of demographic characteristics in Trump Followers on Twitter. They considered a negative binomial regression model for modeling the "likes" and used LDA to extract the tweets of Trump. They provided evaluations on the dataset US2016 (Twitter) that include a number of followers for all the candidates in the United States presidential election of 2016. The authors demonstrated that topic-enhanced word embedding is very impressive for classification of sentiment on Twitter (Wang et al., 2016).

G. Crime prediction/evaluation

Over time; definitely, provides further applications for modeling in various sciences. According to recent work, some researchers have applied the topic modeling methods to crime prediction and analysis. **Chen et al.** introduced an early warning system to find the crime activity intention base on an LDA) model and collaborative representation classifier (CRC).The system includes two steps: They utilized LDA for learning features and extract the features that can represent from article sources. And for the next step, used from achieved features of LDA to classify a new document by collaborative representation classifier (CRC). **Geber et al.** used a statistical topic modeling based on LDA to identify discussion topics among a big city in the United States and used kernel density estimation (KDE) techniques for a standard crime prediction . **Sharma et al.** the authors introduced an approach based on the geographical model of crime intensities to detect the safest path between two locations and used a simple Naive Bayes classifier based on features derived from an LDA model (Chen et al., 2015, Gerber, 2014, Sharma et al., 2015).

Tablel12. Impressive works LDA-based in crime prediction

Study- Author	**Year**	**Purpose**	**Dataset**
(Wang et al., 2012)	2012	Automatic semantic analysis on Twitter posts	A corpus of tweets from Twitter(manual)
(Gerber, 2014)	2014	Crime prediction using tagged tweets	City of Chicago Data: https://data.cityofchicago.or
(Chen et al., 2015)	2015	Detect the crime activity intention	800 news articles from yahoo Chinese news

4. Open source library and tools / datasets / Software packages and tools for the analysis

We need new tools to help us organize, search, and understand these vast amounts of information

4.1 library/tools

Many tools for Topic modeling and analysis are available, including professional and amateur software, commercial software, and open source software and also, there are many popular datasets that can consider as a standard source for testing and evaluation. **Table7**, Show some well-known tools for topic modeling and **Table8**, show some well-known datasets for topic modeling. For example; Mallet tools,The MALLET topic model package incorporates an extremely quick and highly scalable implementation of Gibbs sampling, proficient methods for tools and document-topic hyperparameter optimization for inferring topics for new documents given trained models. Topic models provide a simple approach to analyze huge volumes of unlabeled text. The role of these tools, as mentioned, A "topic" consists of a group of words that habitually happen together. Topic models can associate words with distinguish and similar meanings among uses of words with various meanings and considering contextual clues. (Steyvers and Griffiths, 2007)

Tablel13. Some well-known tools for topic modeling

Tools	Implementation/ Language	Inference/Parameter	source code availability
Mallet (McCallum, 2002)	Java	Gibbs sampling	http://mallet.cs.umass.edu/topics.php
TMT (Ramage and Rosen, 2011)	Java	Gibbs sampling	https://nlp.stanford.edu/software/tmt/tmt-0.4/
Mr.LDA (Zhai et al., 2012)	Java	Variational Bayesian inference	https://github.com/lintool/Mr.LDA
JGibbLDA (Phan and Nguyen, 2006)	Java	Gibbs sampling	http://jgibblda.sourceforge.net/
Gensim (Řehůřek and Sojka, 2011)	Python	Gibbs sampling	https://radimrehurek.com/gensim
TopicXP (Řehůřek and Sojka, 2011)	Java(Eclipse plugin)		http://www.cs.wm.edu/semeru/TopicXP/
Matlab Topic Modeling (Steyvers and Griffiths, 2011)	Matlab	Gibbs sampling	http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
Yahoo_LDA(Chang, 2011)	C++	Gibbsampling	https://github.com/shravanmn/Yahoo_LDA
Lda in R (Ahmed et al., 2012)	R	Gibbsampling	https://cran.r-project.org/web/packages/lda/

For evaluation and testing, according to previous work, researchers have released many dataset in various subjects, size, and dimensions for public access and other future work. So, due to the importance of this research, we examined the well-known dataset from previous work. **Table 8,** shows lists of some famous and popular datasets in various languages.

Tablel14. Some well-known Dataset for topic modeling

Dataset	**Language**	**Date of publish**	**Short-detail**	**Availability address**
Reuters (Reuters21578) (Lewis, 1997)	English	*1997*	Newsletters in various categories	http://kdd.ics.uci.edu/databases/reuters21578/reuters21578
ReutersV 1 (Reuters-Volume I) (Lewis et al., 2004)	English	2004	Newsletters in various categories	
UDI-TwitterCrawl-Aug2012 (Li et al., 2012)	English	2012	-a twitter dataset from millions of tweets	https://wiki.illinois.edu//wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012
SemEval-2013 Dataset (Manandhar and Yuret, 2013)	English	2013	-a twitter dataset from millions of tweets	
Wiki10[179]	English	2009	a Wikipedia Document in various category	http://nlp.uned.es/social-tagging/wiki10+/
Weibo dataset (Zhang et al., 2013)	Chinese	2013	a popular Chinese microblogging network	
Bag of Words[180]	English	2008	a multi dataset(PubMed abstracts, KOS blog, NYTimes news, NIPS full papers, Enron Emails)	https://archive.ics.uci.edu/ml/datasets/Bag+of+Words
CiteUlike (Wang and Blei, 2011)	English	2011	a bibliography sharing service of	http://www.citeulike.org/faq/data.adp

			academic papers	
DBLP Dataset[183] (Lange and Naumann, 2011)	English		a bibliographic database about computer science journals	https://hpi.de/naumann/projects/repeatability/datasets/dblp-dataset.html
HowNet lexicon	Chinese	2000-2013	A Chinese machine-readable dictionary / lexical knowledge	http://www.keenage.com/html/e_index.html
Virastyar , Persian lexicon(Asgari and Chappelier, 2013)	Persian	2013	Persian poems electronic lexica	http://ganjoor.net/ http://www.virastyar.ir/data/
NIPS abstracts	English	2016	The distribution of words in the full text of the NIPS conference (1987 to 2015)	https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015
Ch-wikipedia (Qin et al., 2016) (Cong et al., 2012)	Chinese		A Chinese corpus from Chinese Wikipedia	
Pascal VOC 2007 (Everingham et al., 2008) (Everingham et al., 2010)	English	2007	a dataset of natural images	http://host.robots.ox.ac.uk/pascal/VOC/voc2007/
AFP_ARB corpus(Larkey and Connell, 2001)	Arabic	2001	A collection of newspaper articless in Arabic from Agence France Presse	
20Newsgroups4 corpus[(Rennie, 2008)	English	Jan 2008	Newsletters in various categories	http://qwone.com/~jason/20Newsgroups/
New York Times (NYT)dataset (Sandhaus, 2008)	English	Oct 2008	Newsletters in various categories	

5. **Seven important issues in Challenges and Open research**

There are challenges and discussions that can be considered as future work in topic modeling. According to our studies, some issues require further research, which can be very effective and attractive for the future. In this section, we will discuss seven important issues and we found that the following issues have not been sufficiently solved. These are the gaps in the reviewed work that would prove to be directions for future work.

5.1 Topics Modeling in image processing, Image classification and annotation:

Image classification and annotation are important problems in computer vision, but rarely considered together and need some intelligent approach for classification. For example, an image of a class highway is more likely annotated with words "road" and "traffic", "car " than words "fish " "scuba" and "boat". **Chong et al.** develop a new probabilistic model for jointly modeling the image, its annotations, and its category label. Their model behaves the class label as a global description of the image and behaves annotation terms as local descriptions of parts of the image. Its underlying probabilistic hypotheses naturally integrate these sources of information. They derive an approximate inference and obtain algorithms based on variational ways as well as impressive approximations for annotating and classifying new images and extended supervised topic modeling (sLDA) to classification problems(Chong et al., 2009).

Lienou and et al. focused on the problem of an image semantic interpretation of large satellite images and used a topic modeling, that each word in a document considering as a segment of image and a document is as an image. For evaluation, they performed experiments on panchromatic QuickBird images. **Wick and et al.** They presented an error correction algorithm using topic modeling based on LDA to Optical character recognition (OCR) error correction. This algorithm including two models: a topic model to calculate the word probabilities and an OCR model for obtaining the probability of character errors. **Vaduva and et al.** introduced a semi-automatic approach to latent information retrieval. according to the hierarchical structure from the images. They considered a combined investigation using LDA model and invariant descriptors of image region for a visual scene modeling. **Philbin and et al.** proposed a geometrically consistent latent topic model to detect significant objects, called Latent Dirichlet Allocation (gLDA). and then introduced methods for the effectiveness of calculations a matching graph, that images are the nodes and the edge strength in visual content. The gLDA method is able to group images of a specific object despite large imaging variations and can also pick out different views of a single object. (Lienou et al., 2010, Philbin et al., 2011, Vaduva et al., 2013, Wick et al., 2007).

5.2 Audio, Music information retrieval and processing

According to our knowledge, few research works have been done in music information analysis using topic modeling. For example; **Nakano et al.** The authors focused on estimation and estimation of singing characteristics from signals of audio. This paper introduces a topic modeling to the vocal timbre analysis, that each song is considered as a weighted mixture of multiple topics. In this approach, first extracted features of vocal timbre of polyphonic music and then used an LDA model to estimate merging weights of multiple topics. For evaluation, they applied 36 songs that consist of 12 Japanese singers. **Hu et al.** They proposed a modified version of LDA to process continuous data and audio retrieval. In this model, each audio document includes various latent topics and considered each topic as a Gaussian distribution on the audio feature data. To evaluate the efficiency of their model, used 1214 audio documents in various categories (such as rain, bell, river, laugh, gun, dog and so on) (Hu et al., 2014, Nakano et al., 2014).

5.3. Drug safety evaluation and Approaches to improving it

Understanding safety of drug and performance continue to be critical and challenging issues for academia and also it is an important issue in new drug discovery. Topic modeling holds potential for mining the biological documents and given the importance and magnitude of this issue, researchers can consider it as a future work. **Yu and et al.** investigated the issue of drug-induced acute liver failure (ALF) with considering the role of topic modeling to drug safety evaluation, they explored the LiverTox database for drugs discovery with a capacity to cause ALF. **Yang and et al.** introduced an automatic approach based on keyphrase extraction to detect expressions of consumer health, according to adverse drug reaction (ADRs) in social media. They used an LDA model as a Feature space modeling to build a topic space on the consumer corpus and consumer health expressions mining. **Bisgin and et al.** introduced an 'in silico' framework to drug repositioning guided through a probabilistic graphical model, that defined a drug as a 'document' and a phenotype form a drug as a 'word'. They applied their approach on the SIDER database to estimate the phenome distribution from drugs and identified 908 drugs from SIDER with new capacity indications and demonstrated that the model can be effective for further investigations (Bisgin et al., 2014, Yang and Kiang, 2015, Yu et al., 2014).

5.4. Analysis of comments of famous personalities, social demographics

Public social media and micro-blogging services, most notably Twitter, the people have found a venue to hear and be heard by their peers without an intermediary. As a consequence and helped by the public nature of twitter political scientists now potentially have the means to evaluate and understand the narratives that organically form, decline among and spread the public in a political campaign. For this field we can refer to some impressive recent works, for example; **Wang and et al.** they introduced a framework to derive the topic preferences of Donald Trump's followers on Twitter and used LDA to infer the weighted mixture for each Trump tweet from topics. **Alashri and et al.** employed sentiment analysis, topic modeling, and trends detection through wavelet transform to topics and trends discovery. They extracted 9,700 posts and 12,050,595 comments of five USA presidential candidates (Ted Cruz , Donald Trump, Hillary Clinton, John Kasich and Bernie Sanders) from their official Facebook pages (Alashri et al., 2016, Wang et al., 2016). **Shi and et al.** The authors proposed a novel probabilistic graphical model for the pattern discovery in comments that called MCTA. This model can cope with the language gap and obtain the common semantics with considering various languages from News Reader Comments (such as Chinese and English newreader comments) (Shi et al., 2016). **Hou and et al.** presented a context and co-mention method using knowledge linking method and a topic-level alignment method to build the links between external resources and news from social media. It can also be said that they applied a unified probabilistic model for predict news and relationship discovery within events and topics with considering the background knowledge of users' comments(Hou et al., 2015) .

5.5. Group discovery and topic modeling

Graph mining and social network analysis in large graphs is a challenging problem. Group discovery has many applications, such as understanding the social structure of organizations, uncovering criminal organizations, and modeling large scale social networks in the Internet community. LDA Models can be an efficient method for discovering latent group structure in large networks. **Henderson and et al.** The authors proposed a scalable Bayesian alternative based on LDA and graph to group discovery in a big real-world graph. For evaluation, they collected three datasets from PubMed. **Yu and et al.** introduced a generative approach using a hierarchical Bayes model for group discovery in Social Media Analysis that called Group Latent Anomaly Detection (GLAD) model. This model merged the ideas from both the LDA model and Mixture Membership Stochastic Block (MMSB) model (Henderson and Eliassi-Rad, 2009, Yu et al., 2015a).

5.6. User Behavior Modeling

Social media provides valuable resources to analyze user behaviors and capture user preferences. Since the user generated data (such as users activities, user interests) in social media is a challenge(Diao et al., 2012) (Yin et al., 2014), using topic modeling techniques(such as LDA) can contribute to an important role for the discovery of hidden structures related to user behavior in social media. Although some topic modeling approaches have been proposed in user behavior modeling, there are still many open questions and challenges to be addressed. For example; **Giri et al.** introduced a novel way using an unsupervised topic model for hidden interests discovery of users and analyzing browsing behavior of users in a cellular network that can be very effective for mobile advertisements and online recommendation systems. **Wang et al.** presented a solution framework based on user behavior and synergetic modeling of multi-modal content using a topic analytic for cross media topic analysis and detection of the behavior of users activities on the web. **Yuan et al.** proposed a framework based on a probabilistic topic modeling method to detection of "user interests" and user behavior pattern discovery in the mobile Web usage log. They applied this model on a real dataset in Beijing that include 3 million users (Giri et al., 2014, Wang et al., 2014b, Yuan et al., 2014). **Siersdorfer and et al.** The authors focused on analysis comment rating behavior of users on social medias and gathered more than 10 million user comments from YouTube and Yahoo! News websites. For YouTube, they restricted their analysis on tag annotations for content and employed Latent Dirichlet Allocation (LDA) to obtain term probabilities and each tag of a video defined as a mixture of latent topics. Also, they used a linear support vector machines (SVMs) to detection of comments likely to attract replies(Siersdorfer et al., 2014).

5.7 Visualizing topic models

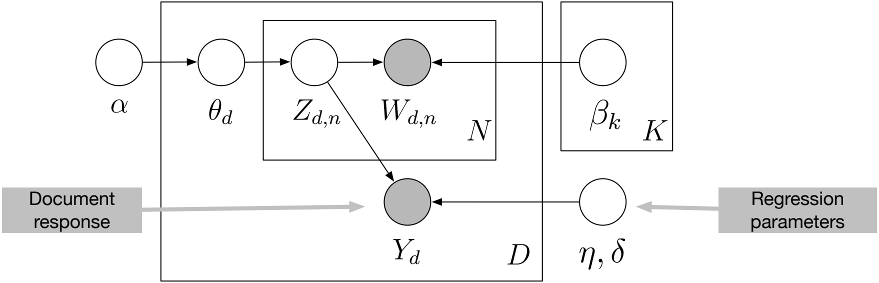
Although different approaches have been investigated to support the visualization of text in large sets of documents such as machine learning, but it is an open challenge in text mining and visualizing data in big data source. Some of the few studies that have been done, such as (Chaney and Blei, 2012, Kim et al., 2017, Murdock and Allen, 2015, Gretarsson et al., 2012). **Chuang and et al.** The authors proposed a topic tool based on a novel visualization technique to the evaluation of textual topical in topic modeling, called Termite. The tool can visualize the collection from the distribution of topic term in LDA with considering a matrix layout. The authors used two measures for understanding a topic model of the Useful terms that including: "saliency" and "distinctiveness". They used the Kullback-Liebler divergence between the topics distribution determined the term for obtain these measures. This tools can increase the interpretations of topical results and make a legible result (Chuang et al., 2012). **Millar and et al.** the authors proposed a combined approach based on Latent Dirichlet Allocation for dimensionality reduction and self-organizing maps to document Clustering and Visualization, that called LDA-SOM [160]. **Sievert and et al.** introduced LDAvis as an interactive visualization system using LDA that capable the providing a global view of the topics, and has a flexible feature for exploring relationship between topics and terms and obtain better understand from a fitted LDA model. It can also be said that the system can find significant topics and cluster them in various categories (Millar et al., 2009, Siersdorfer et al., 2014).

Supervised Topic Models

Supervised LDA

- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
  - User reviews paired with a number of stars
  - Web pages paired with a number of “likes”
  - Documents paired with links to other documents
  - Images paired with a category
- **Supervised LDA** are topic models of documents and responses. They are fit to find topics predictive of the response.

Supervised LDA



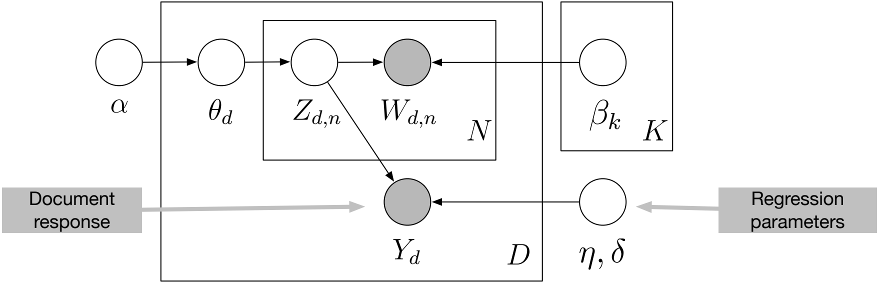
Document response

Regression parameters

1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \textstyle\sum_{n=1}^{N} z_n.$$

Supervised LDA

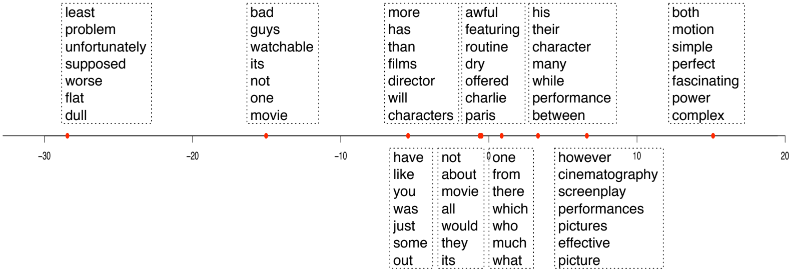


- Fit sLDA parameters to documents and responses.
  This gives: topics $\beta_{1:K}$ and coefficients $\eta_{1:K}$.
- Given a new document, predict its response using the expected value:

$$\mathrm{E}\left[Y \,|\, w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2\right] = \eta^\top \mathrm{E}\left[\bar{Z} \,|\, w_{1:N}\right]$$

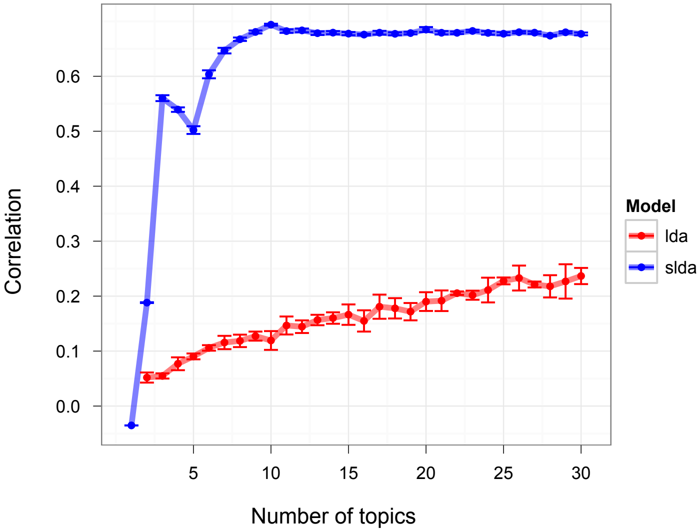
- This blends generative and discriminative modeling.

Supervised LDA

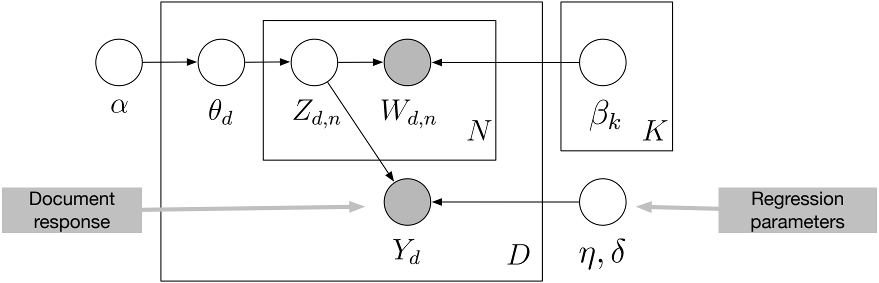


- 10-topic sLDA model on movie reviews (Pang and Lee, 2005).
- Response: number of stars associated with each review
- Each component of coefficient vector η is associated with a topic.

Supervised LDA

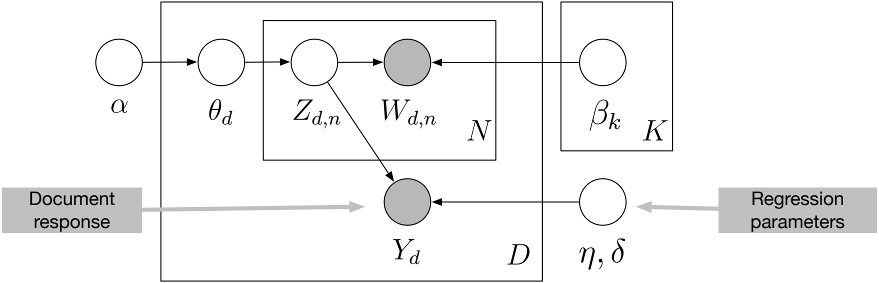


Supervised LDA



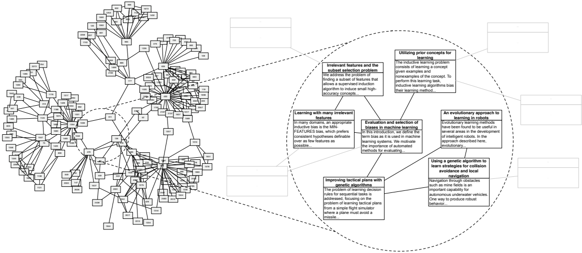
- SLDA enables model-based regression where the predictor is a document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.

Supervised LDA



- SLDA has been extended to generalized linear models, e.g., for image classification and other non-continuous responses.
- We will discuss two extensions of sLDA
  - **Relational topic models**: Models of networks and text
  - **Ideal point topic models**: Models of legislative voting behavior

Relational topic models



Irrelevant features and the subset selection problem
We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...

Utilizing prior concepts for learning
The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...

Learning with many irrelevant features
In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...

Evaluation and selection of biases in machine learning
In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...

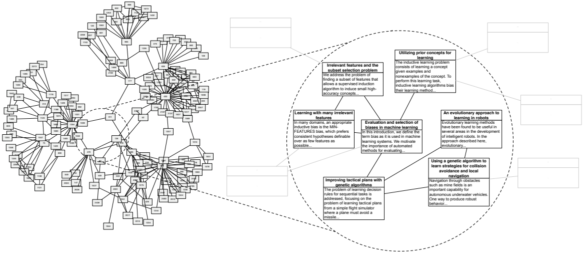
An evolutionary approach to learning in robots
Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...

Using a genetic algorithm to learn strategies for collision avoidance and local navigation
Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...

Improving tactical plans with genetic algorithms
The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

- Many data sets contain **connected observations.**
- For example:
  - Citation networks of documents
  - Hyperlinked networks of web-pages.
  - Friend-connected social network profiles

Relational topic models



Irrelevant features and the subset selection problem
We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...

Utilizing prior concepts for learning
The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...

Learning with many irrelevant features
In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...

Evaluation and selection of biases in machine learning
In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...

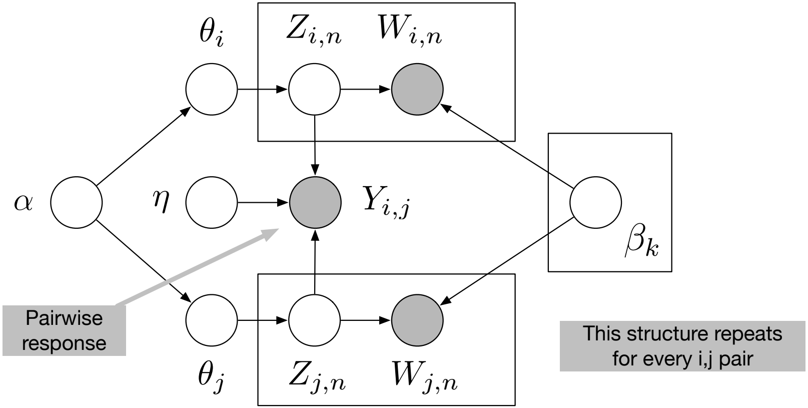
An evolutionary approach to learning in robots
Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...

Using a genetic algorithm to learn strategies for collision avoidance and local navigation
Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...

Improving tactical plans with genetic algorithms
The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

- Research has focused on finding communities and patterns in the link-structure of these networks. But this ignores content.
- We adapted sLDA to pairwise response variables.
  This leads to a model of **content and connection**.
- Relational topic models find related hidden structure in both types of data.

Relational topic models



Pairwise response

This structure repeats for every i,j pair

- Adapt fitting algorithm for sLDA with binary GLM response
- RTMs allow predictions about new and unlinked data.
- These predictions are out of reach for traditional network models.

Relational topic models

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo	RTM (ψ_e)
Rates of convergence of the Hastings and Metropolis algorithms	
Possible biases induced by MCMC convergence diagnostics	
Bounding convergence time of the Gibbs sampler in Bayesian image restoration	
Self regenerative Markov chain Monte Carlo	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation	
Diagnosing convergence of Markov chain Monte Carlo algorithms	
Exact Bound for the Convergence of Metropolis Chains	**LDA + Regression**
Self regenerative Markov chain Monte Carlo	
Minorization conditions and convergence rates for Markov chain Monte Carlo	
Gibbs-markov models	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models	
Mediating instrumental variables	
A qualitative framework for probabilistic inference	
Adaptation for Self Regenerative MCMC	

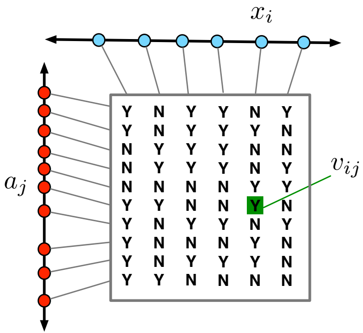
Given a new document, which documents is it likely to link to?

Relational topic models

Competitive environments evolve better solutions for complex tasks	
Coevolving High Level Representations	RTM (ψ_e)
A Survey of Evolutionary Strategies	
Genetic Algorithms in Search, Optimization and Machine Learning	
Strongly typed genetic programming in evolving cooperation strategies	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling...	
Evolutionary Module Acquisition	
An Empirical Investigation of Multi-Parent Recombination Operators...	
A New Algorithm for DNA Sequence Assembly	LDA + Regression
Identification of protein coding regions in genomic DNA	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling...	
A genetic algorithm for passive management	
The Performance of a Genetic Algorithm on a Chaotic Objective Function	
Adaptive global optimization with local search	
Mutation rates as adaptations	

Given a new document, which documents is it likely to link to?

Ideal point topic models

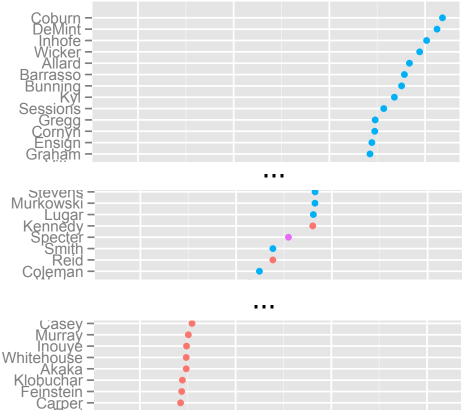


Y	N	Y	Y	N	Y
Y	N	Y	Y	Y	N
N	Y	Y	Y	N	N
N	Y	Y	Y	N	Y
N	N	N	N	Y	Y
Y	Y	N	N	Y	N
Y	N	Y	Y	N	Y
Y	N	N	N	Y	N
Y	N	Y	Y	Y	N
Y	Y	N	N	N	N

$$p(v_{ij}) = f(d(x_i, a_j))$$

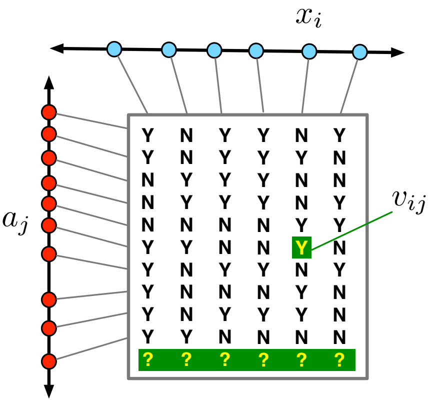
- The **ideal point model** uncovers voting patterns in legislative data
- We observe roll call data v_{ij}.
- Bills attached to discrimination parameters a_j.
  Senators attached to ideal points x_i.

Ideal point topic models



- Posterior inference reveals the political spectrum of senators
- Widely used in quantitative political science.

Ideal point topic models

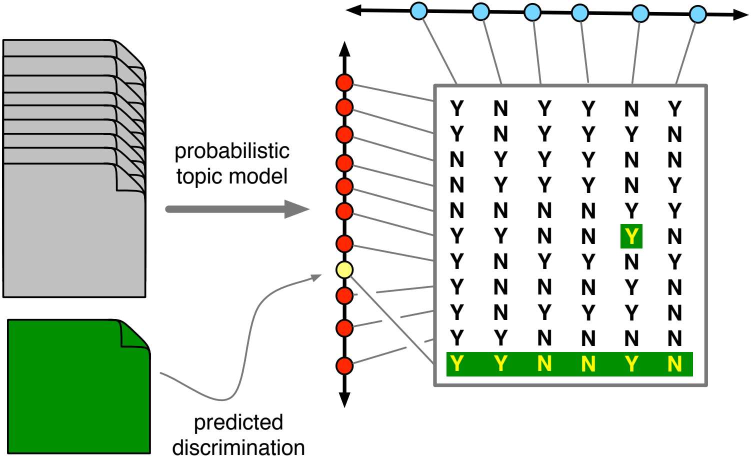


Y	N	Y	Y	N	Y
Y	N	Y	Y	Y	N
N	Y	Y	Y	N	N
N	Y	Y	Y	N	Y
N	N	N	N	Y	Y
Y	Y	N	N	Y	N
Y	N	Y	Y	N	Y
Y	N	N	N	Y	N
Y	N	Y	Y	Y	N
Y	Y	N	N	N	N
?	?	?	?	?	?

$$p(v_{ij}) = f(d(x_i, a_j))$$

- We can predict a missing vote.
- But we cannot predict all the missing votes from a bill.
- Cf. the limitations of collaborative filtering

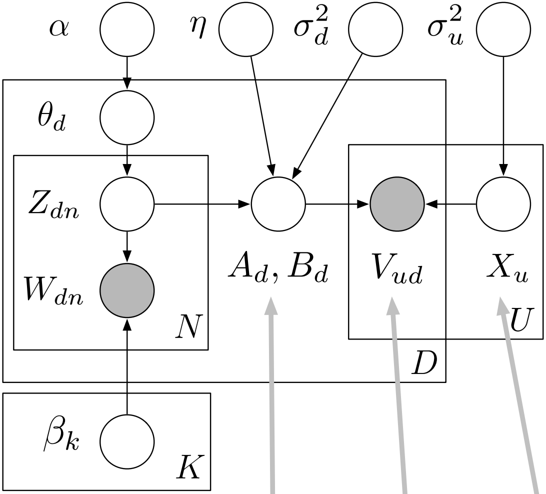
Ideal point topic models



Y	N	Y	Y	N	Y
Y	N	Y	Y	Y	N
N	Y	Y	Y	N	N
N	Y	Y	Y	N	Y
N	N	N	N	Y	Y
Y	Y	N	N	Y	N
Y	N	Y	Y	N	Y
Y	N	N	N	Y	N
Y	N	Y	Y	Y	N
Y	Y	N	N	N	N
Y	Y	N	N	Y	N

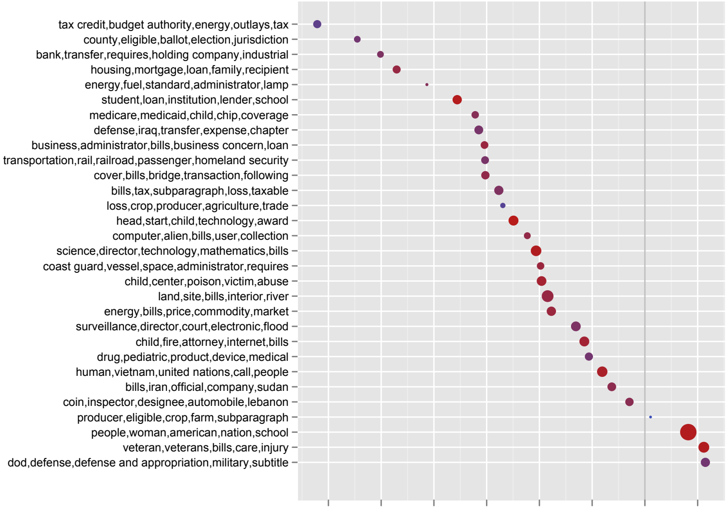
- Use supervised LDA to predict bill discrimination from bill text.
- But this is a **latent response.**

Ideal point topic models



Bill content Bill sentiment Votes Ideal points

Ideal point topic models



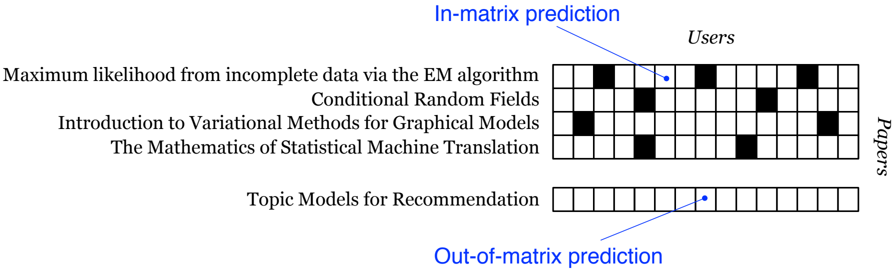
In addition to senators and bills, IPTM places **topics** on the spectrum.

Summary: Supervised topic models

- Many documents are associated with response variables.
- **Supervised LDA** embeds LDA in a generalized linear model that is conditioned on the latent topic assignments.
- **Relational topic models** use sLDA assumptions with pair-wise responses to model networks of documents.
- **Ideal point topic models** demonstrates how the response variables can themselves be latent variables. In this case, they are used downstream in a model of legislative behavior.
- (SLDA, the RTM, and others are implemented in the R package “lda.”)

Modeling User Data and Text

Topic models for recommendation (Wang and Blei, 2011)



In-matrix prediction

Users

Maximum likelihood from incomplete data via the EM algorithm

Conditional Random Fields

Introduction to Variational Methods for Graphical Models

The Mathematics of Statistical Machine Translation

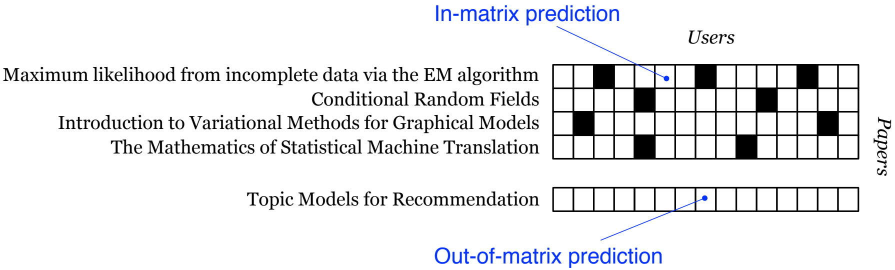
Papers

Topic Models for Recommendation

Out-of-matrix prediction

- In many settings, we have information about **how people use documents.**
- With new models, this can be used to
  - Help people find documents that they are interested in
  - Learn about what the documents mean to the people reading them
  - Learn about the people reading (or voting on) the documents.
- (We also saw this in ideal point topic models.)

Topic models for recommendation (Wang and Blei, 2011)



In-matrix prediction

Users

Maximum likelihood from incomplete data via the EM algorithm

Conditional Random Fields

Introduction to Variational Methods for Graphical Models

The Mathematics of Statistical Machine Translation

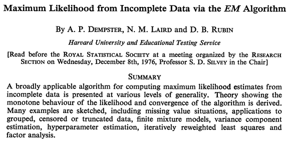
Papers

Topic Models for Recommendation

Out-of-matrix prediction

- Online communities of scientists’ allow for new ways of connecting researchers to the research literature.
- With **collaborative topic models**, we recommend scientific articles based both on other scientists’ preferences and their content.
- We can form both “in-matrix” and “out-of-matrix” predictions. We can learn about which articles are important, and which are interdisciplinary.

- Consider EM (Dempster et al., 1977). The text lets us estimate its topics:



Maximum Likelihood from Incomplete Data via the *EM* Algorithm

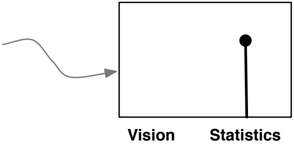
By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

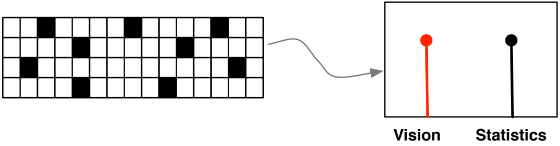
[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

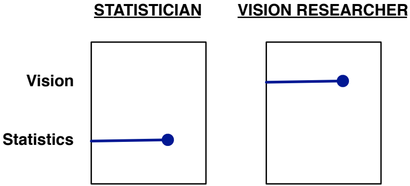
A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.



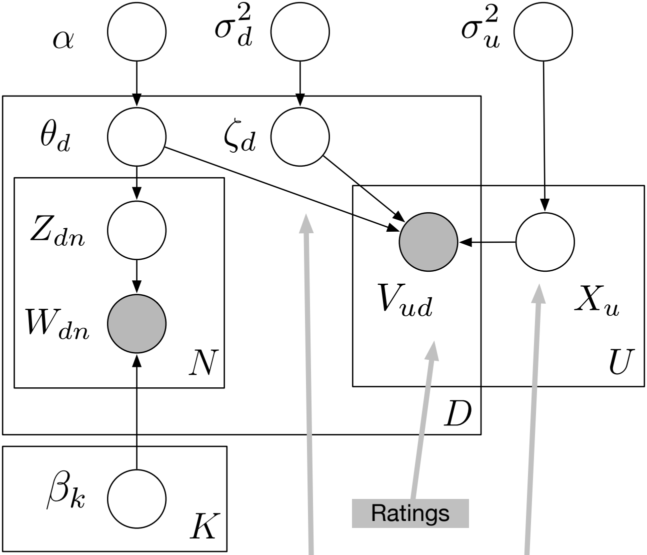
- With user data, we adjust the topics to account for who liked it:



- We can then recommend to users:



Topic models for recommendation



Ratings

Article content

Topic correction

User Preferences

Topic models for recommendation

MENDELEY

- Big data set from Mendeley.com
- Fit the model with **stochastic optimization**
- The data—
  - 261K documents
  - 80K users
  - 10K vocabulary terms
  - 25M observed words
  - 5.1M entries (sparsity is 0.02%)

Maximum Likelihood from Incomplete Data via the *EM* Algorithm

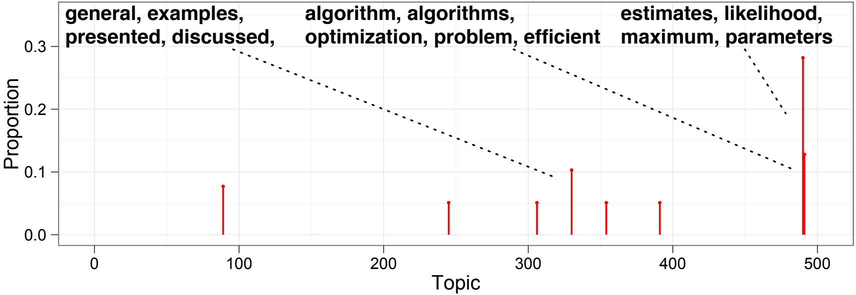
By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

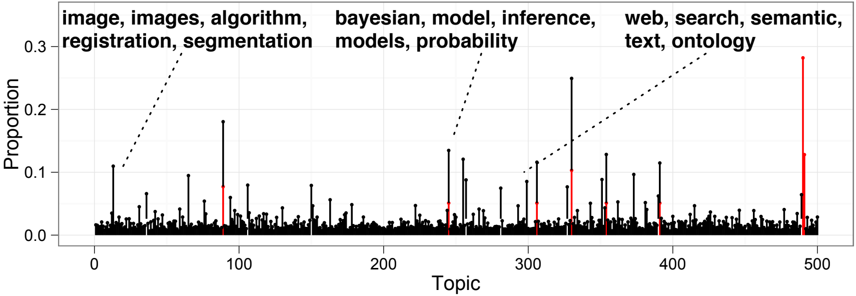
Harvard University and Educational Testing Service

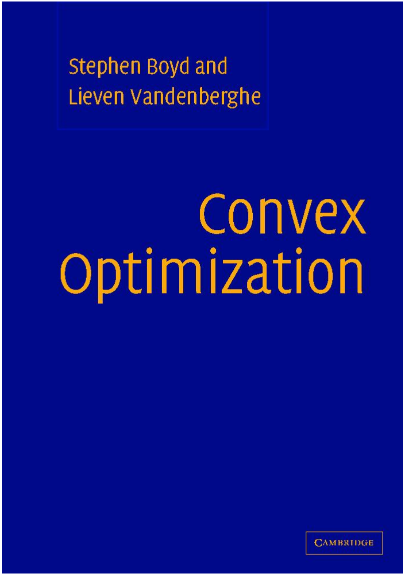
[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. SILVEY in the Chair]

SUMMARY

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.


general, examples,
presented, discussed,
algorithm, algorithms,
optimization, problem, efficient
estimates, likelihood,
maximum, parameters
Proportion
0.0
0.1
0.2
0.3
Topic
0
100
200
300
400
500

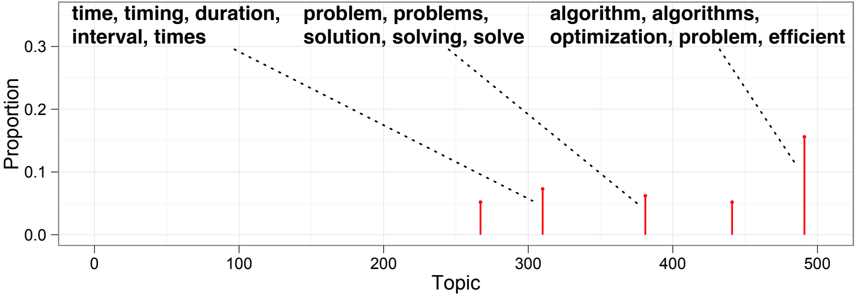

image, images, algorithm, registration, segmentation
bayesian, model, inference, models, probability
web, search, semantic, text, ontology
Proportion
0.0
0.1
0.2
0.3
Topic
0
100
200
300
400
500

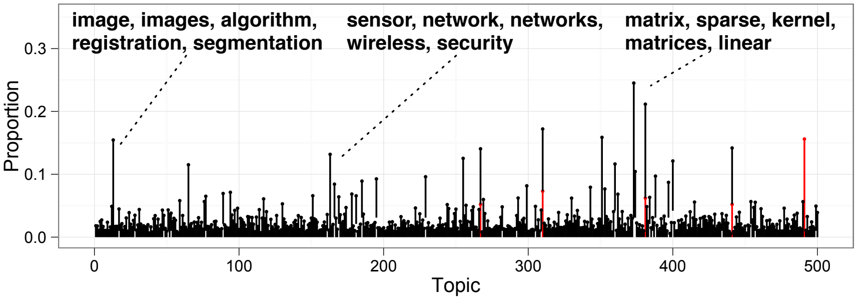


Stephen Boyd and
Lieven Vandenberghe

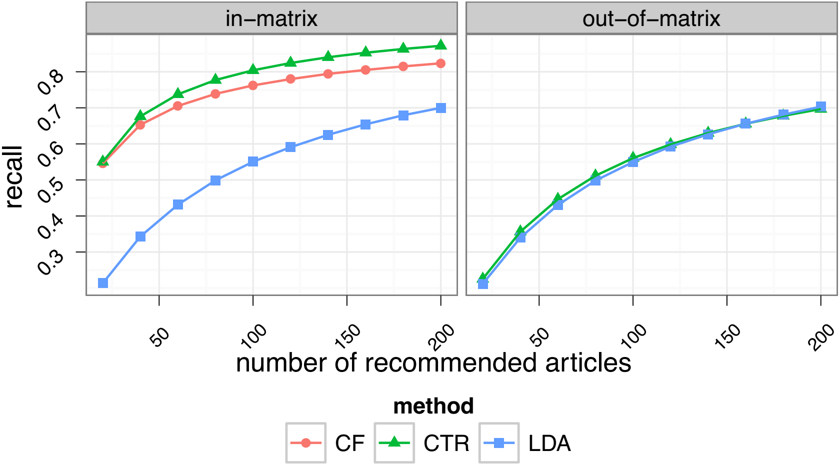
Convex Optimization

CAMBRIDGE


time, timing, duration, interval, times
problem, problems, solution, solving, solve
algorithm, algorithms, optimization, problem, efficient
Proportion
0.0
0.1
0.2
0.3
Topic
0
100
200
300
400
500


image, images, algorithm, registration, segmentation
sensor, network, networks, wireless, security
matrix, sparse, kernel, matrices, linear
Proportion
0.0
0.1
0.2
0.3
Topic
0
100
200
300
400
500

Topic models for recommendation



Can make predictions about current articles and new articles

More than recommendation

Maximum likelihood from incomplete data via the EM algorithm			■					■					■		
Conditional Random Fields					■						■				
Introduction to Variational Methods for Graphical Models		■												■	
The Mathematics of Statistical Machine Translation					■					■					

Papers

- The users also **tell us about the data**.
- We can look at posterior estimates to find
  - Widely read articles in a field
  - Articles in a field that are widely read in other fields
  - Articles from other fields that are widely read in a field
- These kinds of explorations require **interpretable dimensions**. They are not possible with classical matrix factorization.

Maximum Likelihood Estimation

Topic	estimates, likelihood, maximum, parameters, method
In-topic, read in topic	*Maximum Likelihood Estimation of Population Parameters* *Bootstrap Methods: Another Look at the Jackknife* *R. A. Fisher and the Making of Maximum Likelihood*
In-topic, read in other topics	*Maximum Likelihood from Incomplete Data with the EM Algorithm* *Bootstrap Methods: Another Look at the Jackknife* *Tutorial on Maximum Likelihood Estimation*
Out-of-topic, read in topic	*Random Forests* *Identification of Causal Effects Using Instrumental Variables* *Matrix Computations*

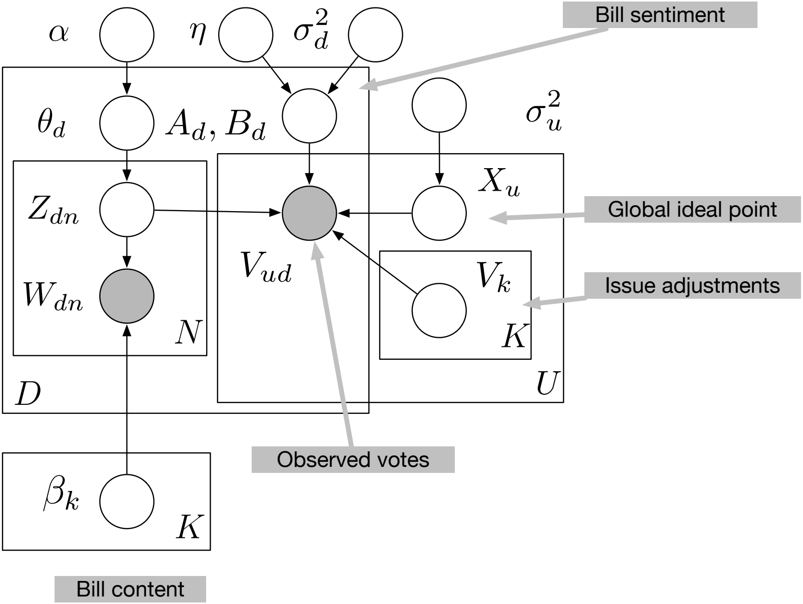
Network Science

Topic	networks, topology, connected, nodes, links, degree
In-topic, read in topic	*Assortative Mixing in Networks* *Characterizing the Dynamical Importance of Network Nodes and Links* *Subgraph Centrality in Complex Networks*
In-topic, read in other topics	*Assortative Mixing in Networks* *The Structure and Function of Complex Networks* *Statistical Mechanics of Complex Networks*
Out-of-topic, read in topic	*Power Law Distributions in Empirical Data* *Graph Structure in the Web* *The Orgins of Bursts and Heavy Tails in Human Dynamics*

Issue-adjusted ideal points

- Our earlier ideal point model uses topics to predict votes from new bills.
- Alternatively, we can use the text to characterize how legislators diverge from their usual ideal points.
- For example: A senator might be left wing, but vote conservatively when it comes to economic matters.

Issue-adjusted ideal points



Bill sentiment

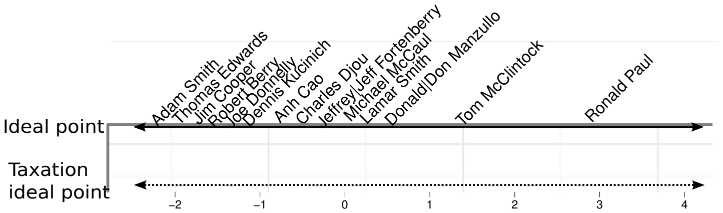
Global ideal point

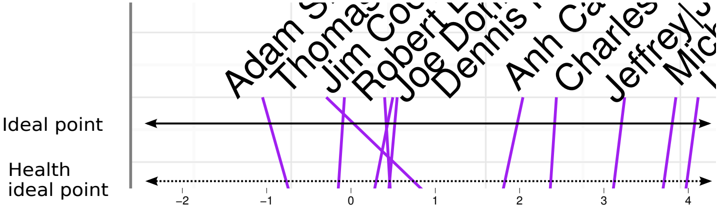
Issue adjustments

Observed votes

Bill content

Issue-adjusted ideal points





Extending LDA

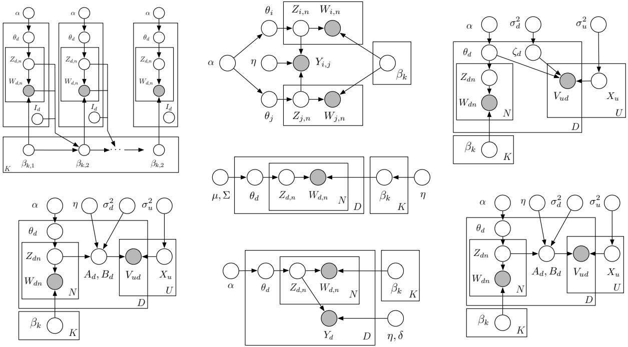
New applications—

- Syntactic topic models
- Topic models on images
- Topic models on social network data
- Topic models on music data
- Topic models for recommendation systems

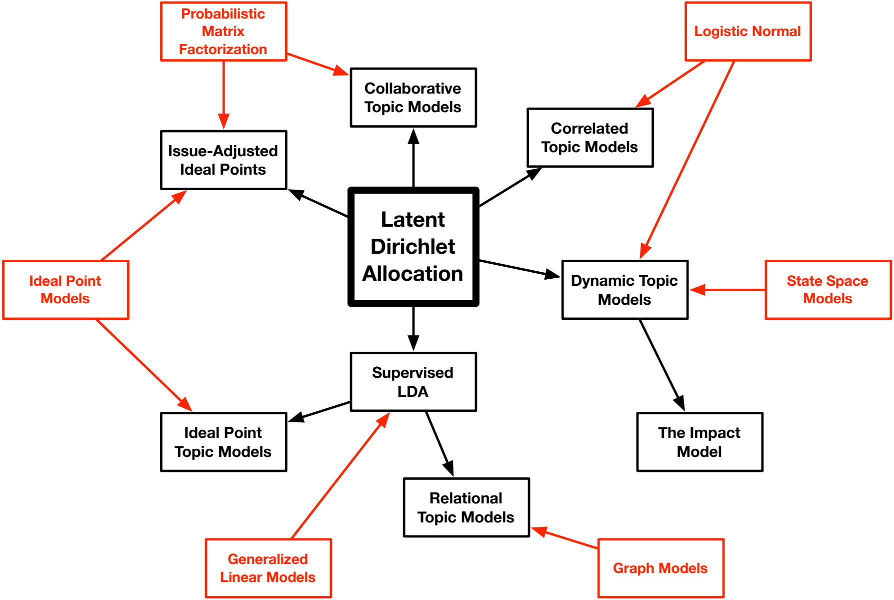
Testing and relaxing assumptions—

- Spike and slab priors
- Models of word contagion
- N-gram topic models

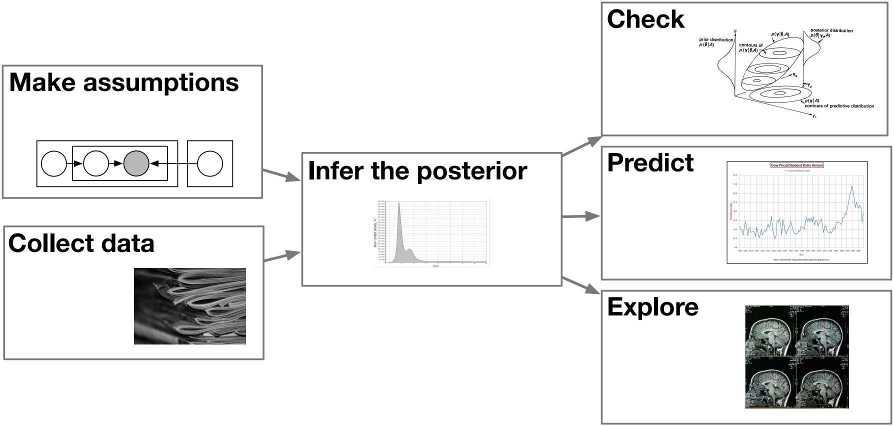
Extending LDA



- Each of these models is tailored to solve a problem.
  - Some problems arise from new kinds of data.
  - Others arise from an issue with existing models.
- Probabilistic modeling is a *flexible and modular language for designing solutions to specific problems*.


Probabilistic Matrix Factorization
Logistic Normal
Collaborative Topic Models
Correlated Topic Models
Issue-Adjusted Ideal Points
Latent Dirichlet Allocation
Ideal Point Models
Dynamic Topic Models
State Space Models
Supervised LDA
Ideal Point Topic Models
The Impact Model
Relational Topic Models
Generalized Linear Models
Graph Models

Extending LDA



Make assumptions

Collect data

Infer the posterior

Check

Predict

Explore

Bayesian Nonparametric Models

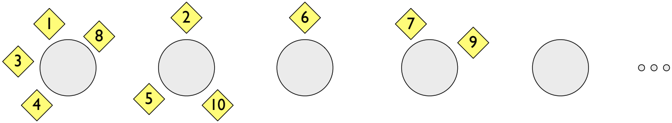
Bayesian nonparametric models

- **Why Bayesian nonparametric models?**
- **The Chinese restaurant process**
- **Chinese restaurant process mixture models**
- **The Chinese restaurant franchise**
- **Bayesian nonparametric topic models**
- **Random measures and stick-breaking constructions**

Why Bayesian nonparametric models?

- Topic models assume that the number of topics is fixed.
- It is a type of **regularization parameter**. It can be determined by cross validation and other model selection techniques.
- Bayesian nonparametric methods skirt model selection—
    - The data determine the number of topics during inference.
    - Future data can exhibit new topics.
- (This is a field unto itself, but has found wide application in topic modeling.)

The Chinese restaurant process (CRP)

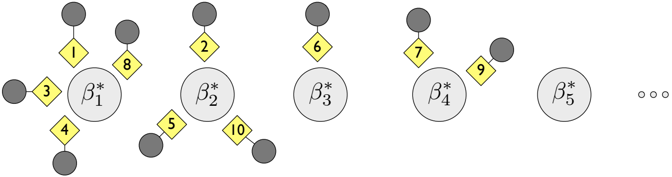


- *N* customers arrive to an infinite-table restaurant. Each sits down according to how many people are sitting at each table,

$$p(z_i = k \mid z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k & \text{for} \quad k \leq K \\ \alpha & \text{for} \quad k = K+1. \end{cases}$$

- The resulting seating plan provides a partition
- This distribution is **exchangeable**: Seating plan probabilities are the same regardless of the order of customers (Pitman, 2002).

CRP mixture models



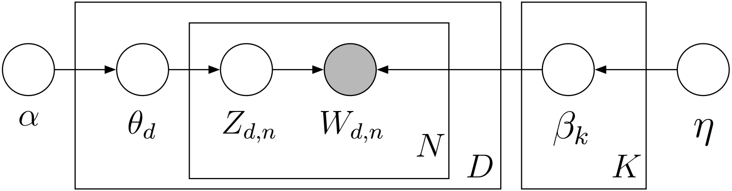
- Associate each table with a topic (β^*).
  Associate each customer with a data point (grey node).

- The number of clusters is infinite a priori;
  the data determines the number of clusters in the posterior.

- Further: the next data point might sit at new table.

- Exchangeability makes inference easy (Escobar and West, 1995; Neal, 2000).

The CRP is not a mixed-membership model

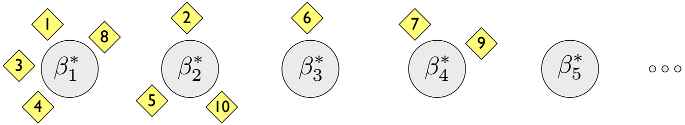


- Mixture models draw each data point from one component.
- The advantage of LDA is that it’s a **mixed-membership model**.
- This is addressed by the **Chinese restaurant franchise**.

The Chinese restaurant franchise (Teh et al., 2006)

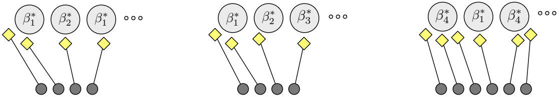
Corpus level restaurant

At the corpus level, topics are drawn from a prior.



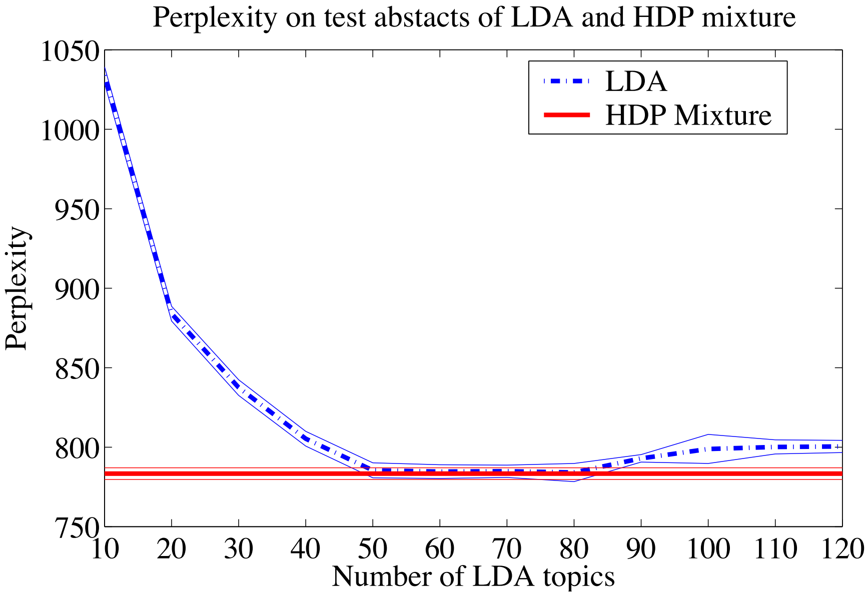
Document level restaurants

Each document-level table is associated with a customer at the corpus level restaurant.

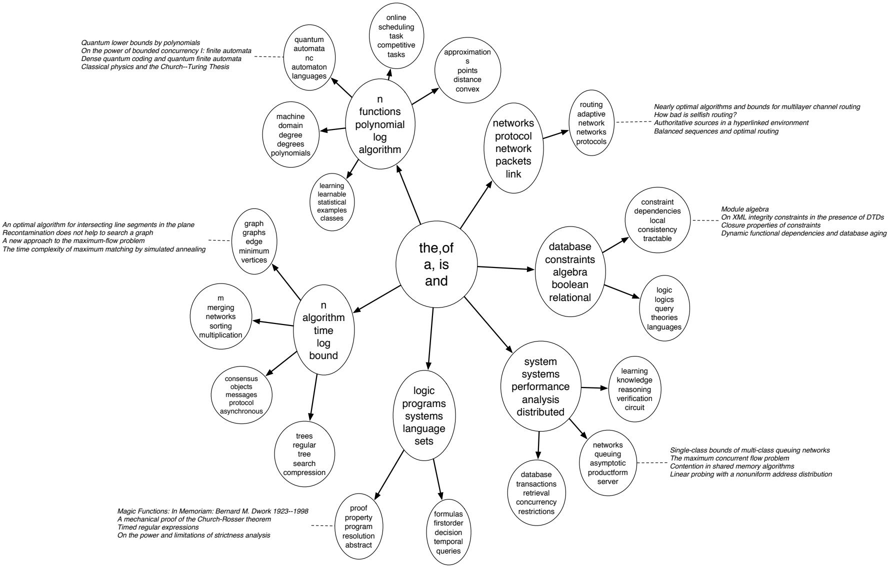


Each word is associated with a customer at the document's restuarant. It is drawn from the topic that its table is associated with.

The CRF selects the “right” number of topics (Teh et al., 2006)



Extended to find hierarchies (Blei et al., 2010)


the,of a, is and
n functions polynomial log algorithm
quantum automata nc automaton languages
online scheduling task competitive tasks
approximations points distance convex
machine domain degree degrees polynomials
learning learnable statistical examples classes
networks protocol network packets link
routing adaptive network networks protocols
database constraints algebra boolean relational
constraint dependencies local consistency tractable
logic logics query theories languages
system systems performance analysis distributed
learning knowledge reasoning verification circuit
networks queuing asymptotic productform server
database transactions retrieval concurrency restrictions
logic programs systems language sets
proof property program resolution abstract
formulas firstorder decision temporal queries
n algorithm time log bound
graph graphs edge minimum vertices
m merging networks sorting multiplication
consensus objects messages protocol asynchronous
trees regular tree search compression

Quantum lower bounds by polynomials
On the power of bounded concurrency I: finite automata
Dense quantum coding and quantum finite automata
Classical physics and the Church--Turing Thesis

Nearly optimal algorithms and bounds for multilayer channel routing
How bad is selfish routing?
Authoritative sources in a hyperlinked environment
Balanced sequences and optimal routing

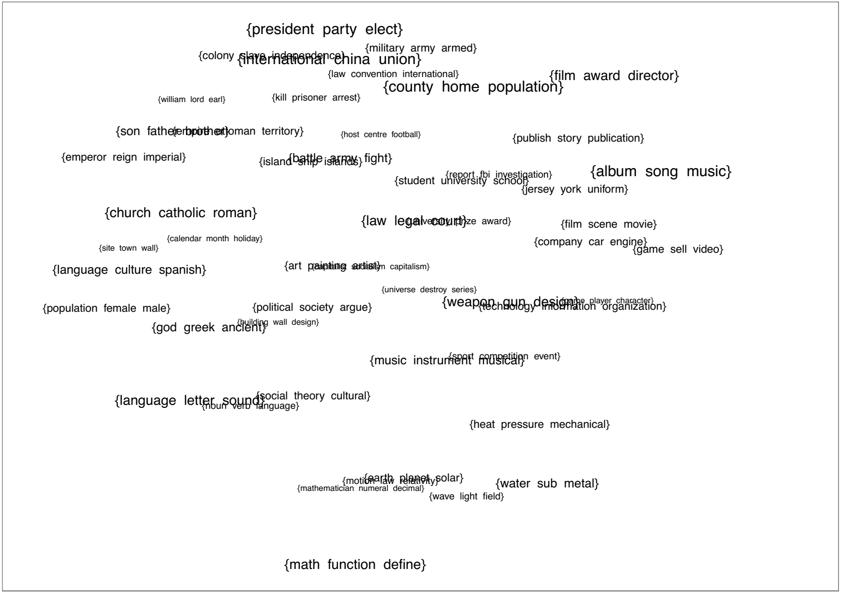
Module algebra
On XML integrity constraints in the presence of DTDs
Closure properties of constraints
Dynamic functional dependencies and database aging

An optimal algorithm for intersecting line segments in the plane
Recontamination does not help to search a graph
A new approach to the maximum-flow problem
The time complexity of maximum matching by simulated annealing

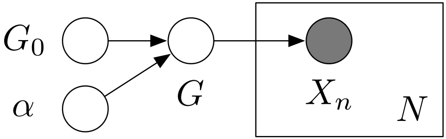
Single-class bounds of multi-class queuing networks
The maximum concurrent flow problem
Contention in shared memory algorithms
Linear probing with a nonuniform address distribution

Magic Functions: In Memoriam: Bernard M. Dwork 1923--1998
A mechanical proof of the Church-Rosser theorem
Timed regular expressions
On the power and limitations of strictness analysis

BNP correlated topic model (Paisley et al., 2011)



Random measures

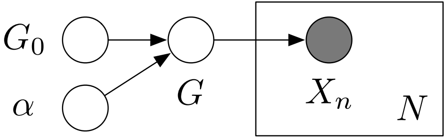


- The CRP metaphors are the best first way to understand BNP methods.
- BNP models were originally developed as **random measure models**.
- E.g., data drawn independently from a random distribution:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha G_0) \\ X_n &\sim G \end{aligned}$$

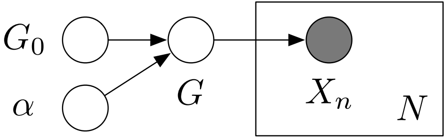
- The random measure perspective helps with certain applications (such as the BNP correlated topic model) and for some approaches to inference.

The Dirichlet process (Ferguson, 1973)



- The Dirichlet process is a distribution of distributions, $G \sim \mathrm{DP}(\alpha, G_0)$
  - *concentration parameter* α (a positive scalar)
  - *base distribution* G_0.
- It produces distributions defined on the same space as its base distribution.

The Dirichlet process (Ferguson, 1973)



- Consider a partition of the probability space $(A_1, \ldots, A_K)$.

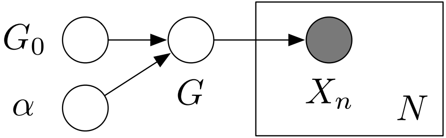
- Ferguson: If for all partitions,

$$\langle G(A_1), \ldots, G(A_k) \rangle \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

  then G is distributed with a Dirichlet process.

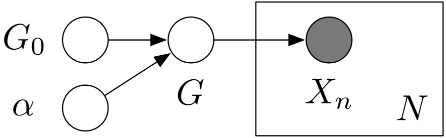
- Note: In this process, the random variables $G(A_k)$ are indexed by the Borel sets of the probability space.

The Dirichlet process (Ferguson, 1973)



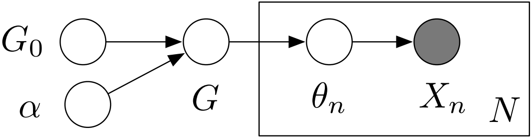
- G is discrete; it places its mass on a countably infinite set of atoms.
- The distribution of the locations is the base distribution G_0.
- As α gets large, G looks more like G_0.
- The conditional $P(G \mid x_{1:N})$ is a Dirichlet process.

The Dirichlet process (Ferguson, 1973)



- Marginalizing out G reveals the **clustering property**.
- The joint distribution of $X_{1:N}$ will exhibit fewer than N unique values.
- These unique values are drawn from G_0.
- The distribution of the partition structure is a $\mathrm{CRP}(\alpha)$.

The Dirichlet process mixture (Antoniak, 1974)

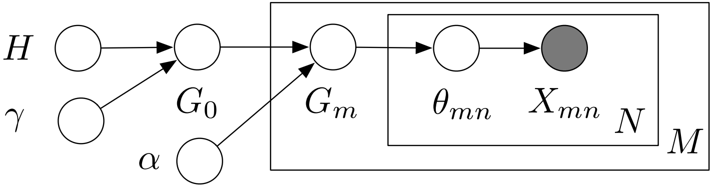


- The draw from G can be a latent parameter to an observed variable:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_n &\sim G \\ x_n &\sim p(\cdot \,|\, \theta_n). \end{aligned}$$

- This smooths the random discrete distribution to a *DP mixture*.
- Because of the clustering property, marginalizing out G reveals that this model is the same as a CRP mixture.

Hierarchical Dirichlet processes (Teh et al., 2006)

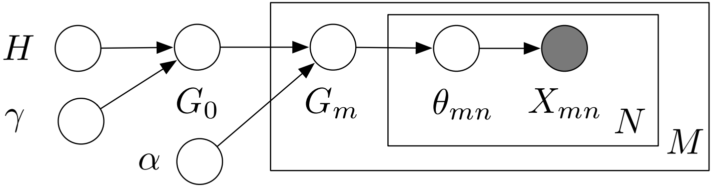


- The hierarchical Dirichlet process (HDP) models *grouped data*.

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_m &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_{mn} &\sim G_m \\ x_{mn} &\sim p(\cdot|\theta_{mn}) \end{aligned}$$

- Marginalizing out G_0 and G_m reveals the Chinese restaurant franchise.

Hierarchical Dirichlet processes (Teh et al., 2006)



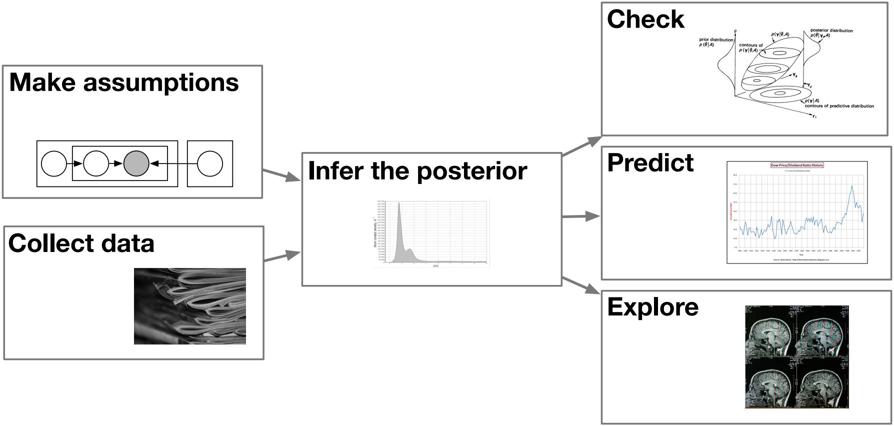
- In topic modeling—
  - The atoms of G_0 are all the topics.
  - Each G_m is a document-specific distribution over those topics
  - The variable θ_{mn} is a topic drawn from G_m.
  - The observation x_{mn} is a word drawn from the topic θ_{mn}.
- Note that in the original topic modeling story, we worked with pointers to topics. Here the θ_{mn} variables are distributions over words.

Summary: Bayesian nonparametrics

- Bayesian nonparametric modeling is a growing field (Hjort et al., 2011).
- BNP methods can define priors over latent combinatorial structures.
- In the posterior, the documents determine the particular form of the structure that is best for the corpus at hand.
- *Recent innovations:*
  - Improved inference (Blei and Jordan, 2006, Wang et al. 2011)
  - BNP models for language (Teh, 2006; Goldwater et al., 2011)
  - Dependent models, such as time series models (MacEachern 1999, Dunson 2010, Blei and Frazier 2011)
  - Predictive models (Hannah et al. 2011)
  - Factorization models (Griffiths and Ghahramani, 2011)

Posterior Inference

Posterior inference



Make assumptions

Collect data

Infer the posterior

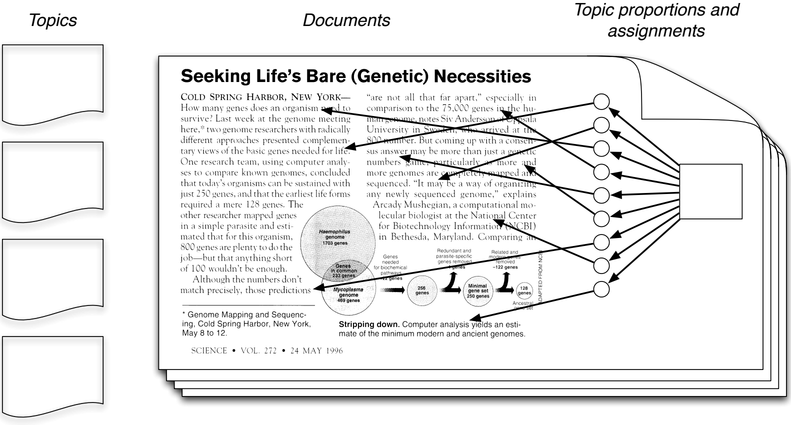
Check

Predict

Explore

- We can express many kinds of assumptions.
- How can we analyze the collection under those assumptions?

Posterior inference


Topics
Documents
Topic proportions and assignments
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways –22 genes
256 genes
Redundant and parasite-specific genes removed –4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

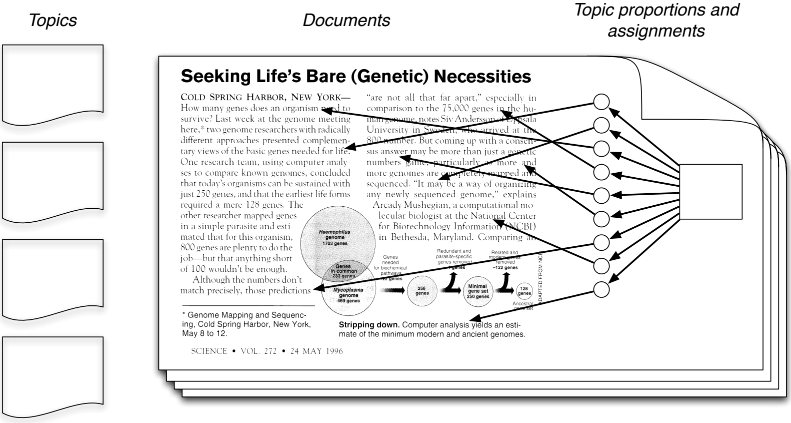
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Posterior inference is the main computational problem.
- Inference links observed data to statistical assumptions.
- Inference on large data is crucial for topic modeling applications.

Posterior inference



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed –6 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

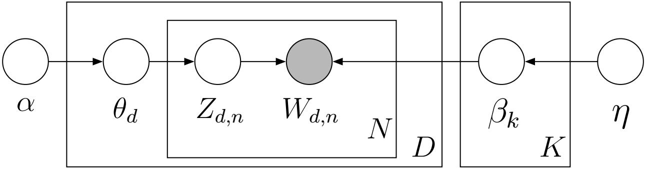
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Our goal is to compute the distribution of the hidden variables conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Posterior inference for LDA



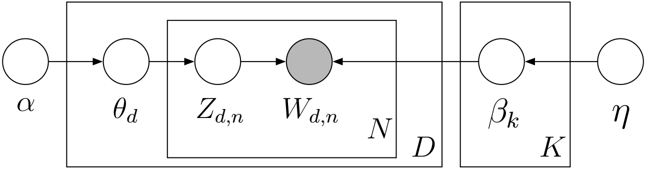
- The joint distribution of the latent variables and documents is

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right).$$

- The posterior of the latent variables given the documents is

$$p(\beta, \theta, \mathbf{z} \mid \mathbf{w}).$$

Posterior inference for LDA

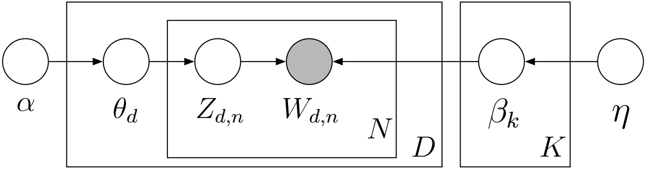


- This is equal to

$$\frac{p(\beta, \theta, \mathbf{z}, \mathbf{w})}{\int_\beta \int_\theta \sum_{\mathbf{z}} p(\beta, \theta, \mathbf{z}, \mathbf{w})}.$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$.
- This is the crux of the inference problem.

Posterior inference for LDA



- There is a large literature on approximating the posterior, both within topic modeling and Bayesian statistics in general.
- We will focus on **mean-field variational methods**.
- We will derive **stochastic variational inference**, a generic approximate inference method for very large data sets.

Variational inference

- Variational inference turns posterior inference into **optimization**.
- The main idea—
  - Place a distribution over the hidden variables with free parameters, called **variational parameters**.
  - Optimize the variational parameters to make the distribution close (in KL divergence) to the true posterior
- Variational inference can be faster than sampling-based approaches.
- It is easier to handle **nonconjugate** models with variational inference. (This is important in the CTM, DTM, and legislative models.)
- It can be scaled up to very large data sets with **stochastic optimization**.

Stochastic variational inference

- We want to condition on large data sets and approximate the posterior.
- In **variational inference**, we optimize over a family of distributions to find the member closest in KL divergence to the posterior.
- Variational inference usually results in an algorithm like this:
  - Infer local variables for each data point.
  - Based on these local inferences, re-infer global variables.
  - Repeat.

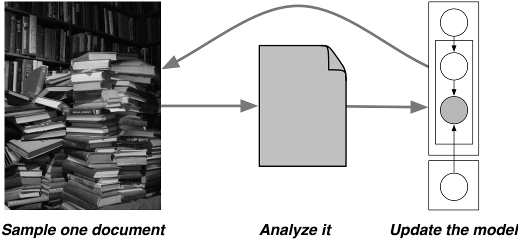
Stochastic variational inference

- This is inefficient. We should know something about the global structure after seeing part of the data.
- And, it assumes a finite amount of data. We want algorithms that can handle **data sources**, information arriving in a constant stream.
- With **stochastic variational inference**, we can condition on large data and approximate the posterior of complex models.

Stochastic variational inference

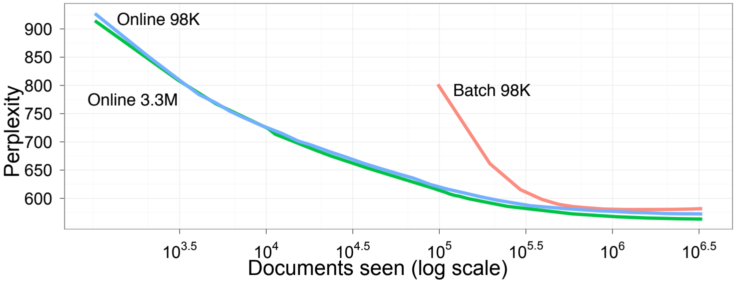
- The structure of the algorithm is:
  - Subsample the data—one data point or a small batch.
  - Infer local variables for the subsample.
  - Update the current estimate of the posterior of the global variables.
  - Repeat.
- This is **efficient**—we need only process one data point at a time.
- We will show: Just as easy as “classical” variational inference

Stochastic variational inference for LDA



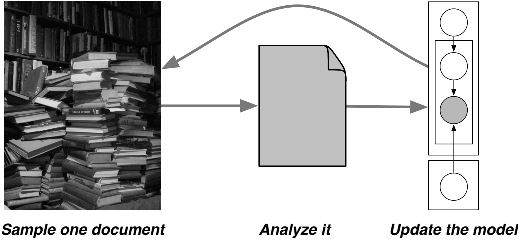
1. Sample a document w_d from the collection
2. Infer how w_d exhibits the current topics
3. Create intermediate topics, formed as though the w_d is the only document.
4. Adjust the current topics according to the intermediate topics.
5. Repeat.

Stochastic variational inference for LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

Stochastic variational inference for LDA



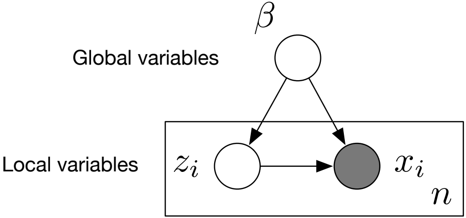
We have developed stochastic variational inference algorithms for

- Latent Dirichlet allocation
- The hierarchical Dirichlet process
- The discrete infinite logistic normal
- Mixed-membership stochastic blockmodels
- Bayesian nonparametric factor analysis
- Recommendation models and legislative models

Organization

- Describe a generic class of models
- Derive mean-field variational inference in this class
- Derive natural gradients for the variational objective
- Review stochastic optimization
- Derive stochastic variational inference

Organization

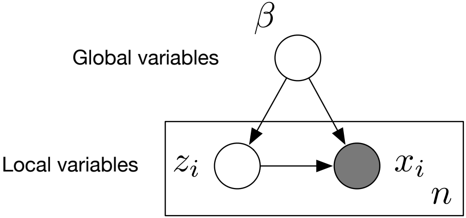


- We consider a **generic model**.
    - Hidden variables are local or global.

- We use **variational inference**.
    - Optimize a simple proxy distribution to be close to the posterior
    - Closeness is measured with Kullback-Leibler divergence

- Solve the optimization problem with **stochastic optimization.**
    - Stochastic gradients are formed by subsampling from the data.

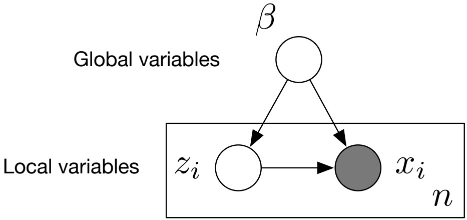
Generic model



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- The observations are $x = x_{1:n}$.
- The **local** variables are $z = z_{1:n}$.
- Th **global** variables are β.
- The *i*th data point x_i only depends on z_i and β.
- Our goal is to compute $p(\beta, z \mid x)$.

Generic model



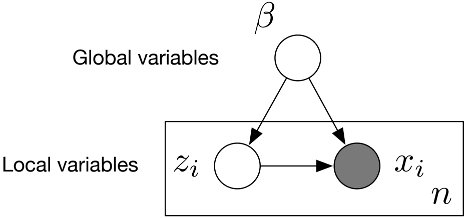
$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variable.

- Assume each complete conditional is in the exponential family,

$$\begin{aligned} p(z_i \mid \beta, x_i) &= h(z_i) \exp\{\eta_\ell(\beta, x_i)^\top z_i - a(\eta_\ell(\beta, x_i))\} \\ p(\beta \mid z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\}. \end{aligned}$$

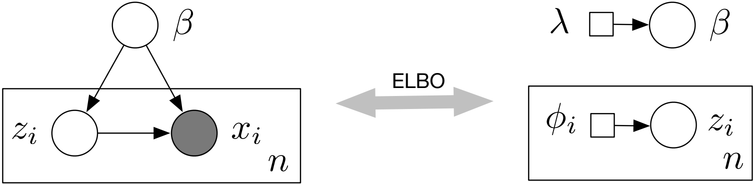
Generic model



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- Bayesian mixture models
- Time series models (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models (LDA and some variants)

Mean-field variational inference



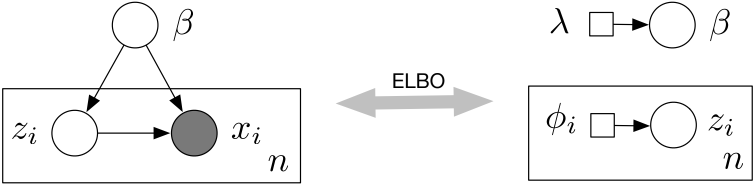
- Introduce a **variational distribution** over the latent variables $q(\beta, z)$.

- We optimize the **evidence lower bound** (ELBO) with respect to q,

$$\log p(x) \geq \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].$$

- Up to a constant, this is the negative KL between q and the posterior.

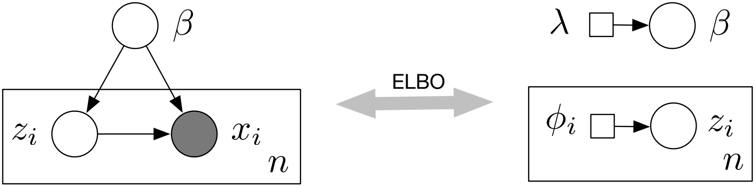
Mean-field variational inference



We can derive the ELBO with Jensen's inequality:

$$\begin{aligned}
\log p(x) &= \log \int p(\beta, Z, X) dZ d\beta \\
&= \log \int p(\beta, Z, X) \frac{q(\beta, Z)}{q(\beta, Z)} dZ d\beta \\
&\geq \int q(\beta, Z) \log \frac{p(\beta, Z, X)}{q(Z)} dZ d\beta \\
&= \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].
\end{aligned}$$

Mean-field variational inference



- We specify $q(\beta, z)$ to be a fully factored variational distribution,

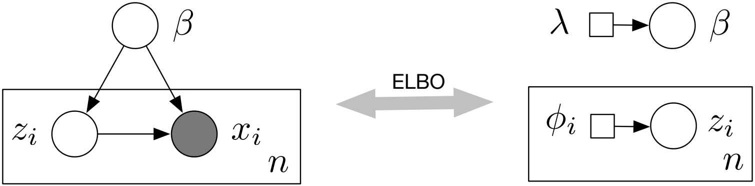
$$q(\beta, z) = q(\beta \,|\, \lambda) \prod_{i=1}^{n} q(z_i \,|\, \phi_i).$$

- Each instance of each variable has its own distribution.
- Each component is in the same family as the model conditional,

$$\begin{aligned} p(\beta \,|\, z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\} \\ q(\beta \,|\, \lambda) &= h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\} \end{aligned}$$

(And, same for the local variational parameters.)

Mean-field variational inference

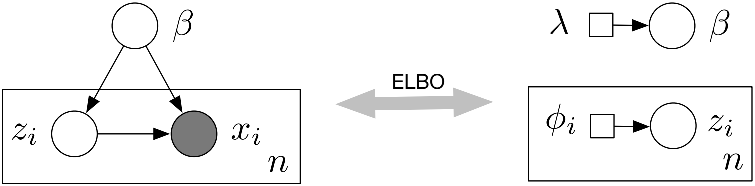


- We optimize the ELBO with respect to these parameters,

$$\mathscr{L}(\lambda, \phi_{1:n}) = \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].$$

- Same as finding the $q(\beta, z)$ that is closest in KL divergence to $p(\beta, z \mid x)$
- The ELBO links the observations/model to the variational distribution.

Mean-field variational inference



- Coordinate ascent: Iteratively update each parameter, holding others fixed.

- With respect to the global parameter, the gradient is

$$\nabla_\lambda \mathscr{L} = a''(\lambda)(\mathrm{E}_\phi[\eta_g(Z,x)] - \lambda).$$

  This leads to a simple coordinate update

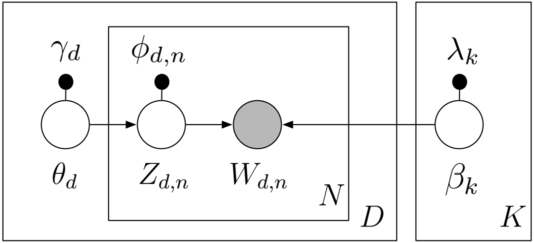
$$\lambda^* = \mathrm{E}_\phi\left[\eta_g(Z,x)\right].$$

- The local parameter is analogous.

Mean-field variational inference

```
Initialize $\lambda$ randomly.
Repeat until the ELBO converges
    1. For each data point, update the local variational parameters:
           $\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$
    2. Update the global variational parameters:
           $\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$
```

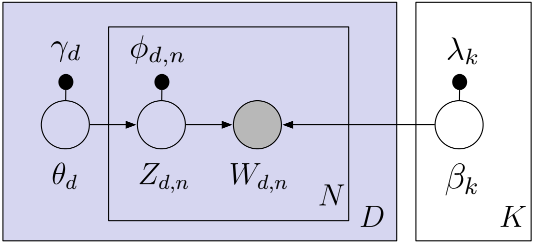
Mean-field variational inference for LDA



- Document variables: Topic proportions θ and topic assignments $z_{1:N}$.
- Corpus variables: Topics $\beta_{1:K}$
- The variational distribution is

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k \,|\, \lambda_k) \prod_{d=1}^{D} q(\theta_d \,|\, \gamma_d) \prod_{n=1}^{N} q(z_{d,n} \,|\, \phi_{d,n})$$

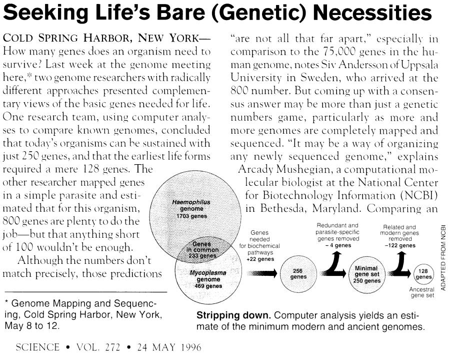
Mean-field variational inference for LDA



- In the "local step" we iteratively update the parameters for each document, holding the topic parameters fixed.

$$\begin{aligned}\gamma^{(t+1)} &= \alpha + \textstyle\sum_{n=1}^{N} \phi_n^{(t)} \\ \phi_n^{(t+1)} &\propto \exp\{\mathbb{E}_q[\log \theta] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}.\end{aligned}$$

Mean-field variational inference for LDA



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

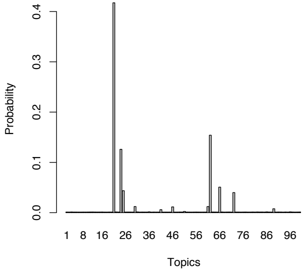
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

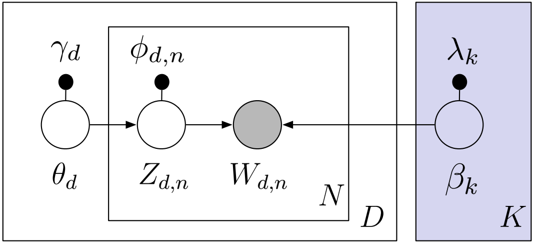
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



Mean-field variational inference for LDA



- In the “global step” we aggregate the parameters computed from the local step and update the parameters for the topics,

$$\lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}.$$

Mean-field variational inference for LDA

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Mean-field variational inference for LDA

```
Initialize topics randomly.
repeat
    for each document do
        repeat
            Update the topic assignment variational parameters.
            Update the topic proportions variational parameters.
        until document objective converges
    end for
    Update the topics from aggregated per-document parameters.
until corpus objective converges.
```

Mean-field variational inference

```
Initialize $\lambda$ randomly.
Repeat until the ELBO converges
    1 Update the local variational parameters for each data point,
        $\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$
    2 Update the global variational parameters,
        $\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$
```

- Note the relationship to existing algorithms like EM and Gibbs sampling.
- But we must analyze the whole data set before completing one iteration.

Mean-field variational inference

Initialize λ randomly.

Repeat until the ELBO converges

1. Update the local variational parameters for each data point,

$$\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$$

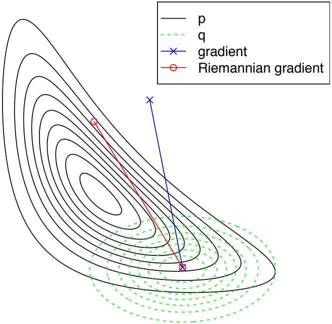
2. Update the global variational parameters,

$$\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

To make this more efficient, we need two ideas:

- Natural gradients
- Stochastic optimization

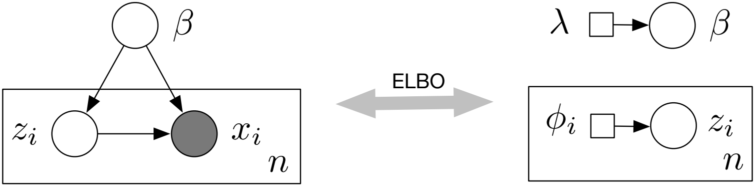
The natural gradient



(from Honkela et al., 2010)

- In natural gradient ascent, we premultiply the gradient by the inverse of a Riemannian metric. Amari (1998) showed this is the steepest direction.
- For distributions, the Riemannian metric is the Fisher information.

The natural gradient



- In the exponential family, the Fisher information is the second derivative of the log normalizer,

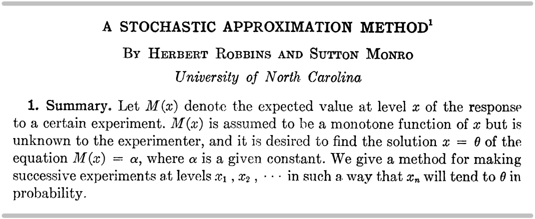
$$G = a''(\lambda).$$

- So, the natural gradient of the ELBO is

$$\hat{\nabla}_\lambda \mathscr{L} = \mathrm{E}_\phi[\eta_g(Z,x)] - \lambda.$$

- We can compute the natural gradient by computing the coordinate updates in parallel and subtracting the current variational parameters.

Stochastic optimization



A STOCHASTIC APPROXIMATION METHOD[1]

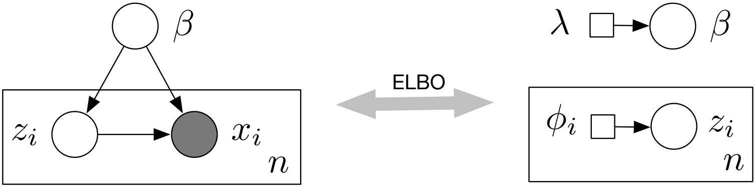
BY HERBERT ROBBINS AND SUTTON MONRO

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

- Why waste time with the real gradient, when a cheaper noisy estimate of the gradient will do (Robbins and Monro, 1951)?
- Idea: Follow a noisy estimate of the gradient with a step-size.
- By decreasing the step-size according to a certain schedule, we guarantee convergence to a local optimum.

Stochastic optimization



- We will use stochastic optimization for global variables.
- Let $\nabla_\lambda \mathscr{L}_t$ be a realization of a random variable whose expectation is $\nabla_\lambda \mathscr{L}$.
- Iteratively set

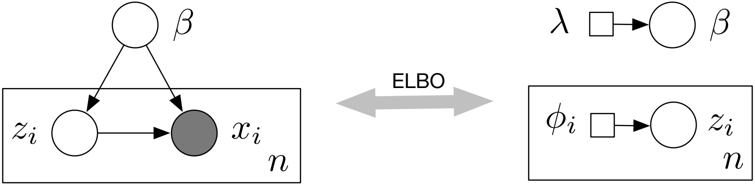
$$\lambda^{(t)} = \lambda^{(t-1)} + \epsilon_t \nabla_\lambda \mathscr{L}_t$$

- This leads to a local optimum when

$$\begin{array}{rcl} \sum_{t=1}^{\infty} \epsilon_t & = & \infty \\ \sum_{t=1}^{\infty} \epsilon_t^2 & < & \infty \end{array}$$

- Next step: Form a noisy gradient.

A noisy natural gradient



- We need to look more closely at the conditional distribution of the global hidden variable given the local hidden variables and observations.

- The form of the local joint distribution is

$$p(z_i, x_i \,|\, \beta) = h(z_i, x_i)\exp\{\beta^\top f(z_i, x_i) - a(\beta)\}.$$

  This means the conditional parameter of β is

$$\eta_g(z_{1:n}, x_{1:n}) = \langle \alpha_1 + \textstyle\sum_{i=1}^{n} f(z_i, x_i), \alpha_2 + n \rangle.$$

- See the discussion of conjugacy in Bernardo and Smith (1994).

A noisy natural gradient

- With local and global variables, we decompose the ELBO

$$\mathscr{L} = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + \textstyle\sum_{i=1}^{n} \mathrm{E}[\log p(z_i, x_i \,|\, \beta)] - \mathrm{E}[\log q(z_i)]$$

- Sample a single data point t uniformly from the data and define

$$\mathscr{L}_t = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + n(\mathrm{E}[\log p(z_t, x_t \,|\, \beta)] - \mathrm{E}[\log q(z_t)]).$$

1. **The ELBO is the expectation of $\mathscr{L}_t$ with respect to the sample.**
2. **The gradient of the t-ELBO is a noisy gradient of the ELBO.**
3. **The t-ELBO is like an ELBO where we saw x_t repeatedly.**

A noisy natural gradient

- Define the conditional as though our whole data set is *n* replications of x_t,

$$\eta_t(z_t, x_t) = \langle \alpha_1 + n \cdot f(z_t, x_t), \alpha_2 + n \rangle$$

- The noisy natural gradient of the ELBO is

$$\nabla_\lambda \hat{\mathscr{L}}_t = \mathrm{E}_{\phi_t}[\eta_t(Z_t, x_t)] - \lambda.$$

- This only requires the local variational parameters of one data point.

- In contrast, the full natural gradient requires all local parameters.

Stochastic variational inference

Initialize global parameters λ randomly.

Set the step-size schedule ϵ_t appropriately.

Repeat forever

1. Sample a data point uniformly,

$$x_t \sim \mathrm{Uniform}(x_1, \ldots, x_n).$$

2. Compute its local variational parameter,

$$\phi = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_t)].$$

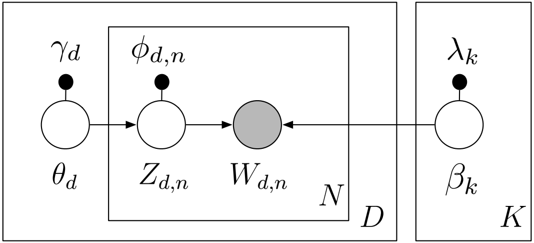
3. Pretend its the only data point in the data set,

$$\hat{\lambda} = \mathrm{E}_{\phi}[\eta_t(Z_t, x_t)].$$

4. Update the current global variational parameter,

$$\lambda^{(t)} = (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \hat{\lambda}.$$

Stochastic variational inference in LDA

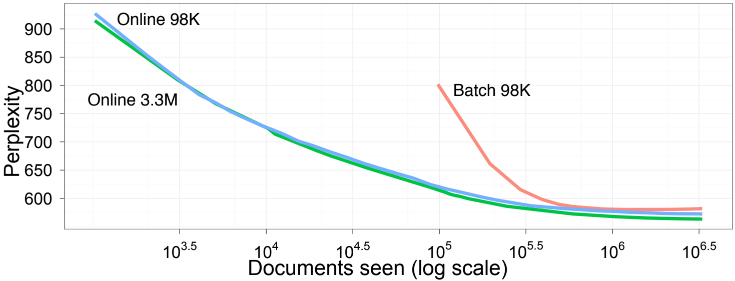


1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form “fake topics” from those local parameters
4. Update the topics to be a weighted average of “fake” and current topics

Stochastic variational inference in LDA

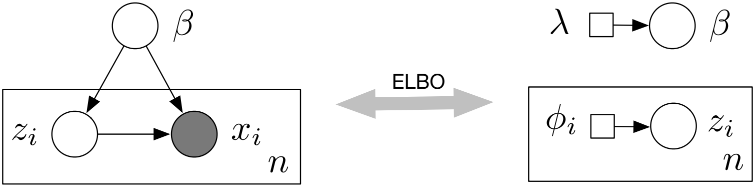
```
Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
**for** $t = 0$ to $\infty$ **do**
    Choose a random document $w_t$
    Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
    **repeat**
        Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$
        Set $\gamma_t = \alpha + \sum_n \phi_{t,n}$
    **until** $\frac{1}{K} \sum_k |\text{change in } \gamma_{t,k}| < \epsilon$
    Compute $\tilde{\lambda}_k = \eta + D \sum_n w_{t,n} \phi_{t,n}$
    Set $\lambda_k = (1 - \rho_t)\lambda_k + \rho_t \tilde{\lambda}_k$.
**end for**
```

Stochastic variational inference in LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems road made service announced national west language	systems health communication service billion language care road	service systems health companies market communication company billion	service systems companies business company billion health industry	service companies systems business company industry market billion	business service companies industry company management systems services	business service companies industry services company management public	business industry service companies services company management public

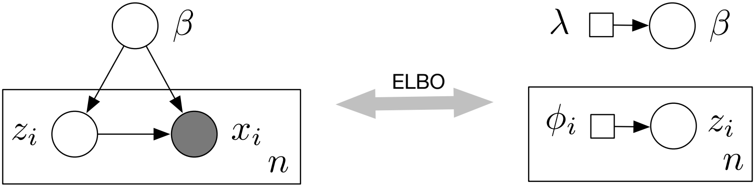
Stochastic variational inference



We defined a generic algorithm for scalable variational inference.

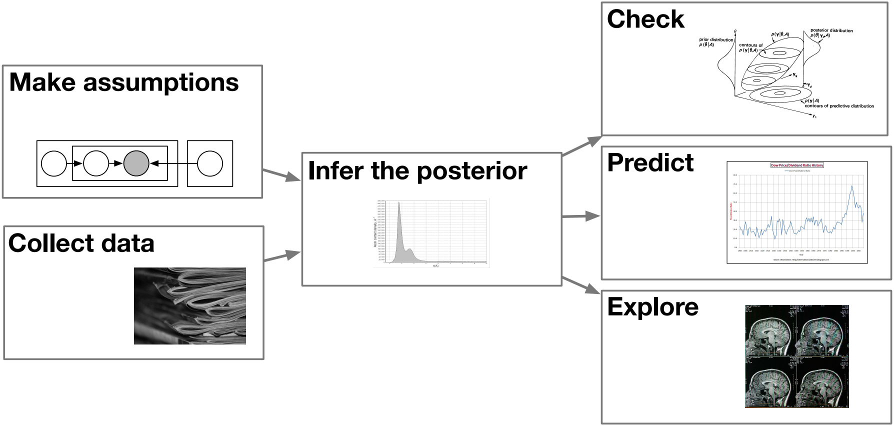
- Bayesian mixture models
- Time series models
  (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization
  (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression
  (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models
  (LDA and some variants)

Stochastic variational inference



- See Hoffman et al. (2010) for LDA (and code).
- See Wang et al. (2010) for Bayesian nonparametric models (and code).
- See Sato (2001) for the original stochastic variational inference.
- See Honkela et al. (2010) for natural gradients and variational inference.

Stochastic variational inference



Make assumptions

Collect data

Infer the posterior

Check

Predict

Explore

prior distribution $p(\theta|A)$
contours of $p(\mathbf{y}|\theta,A)$
$p(\mathbf{y}|\theta,A)$
posterior distribution $p(\theta|\mathbf{y}_d,A)$
$p(\mathbf{y}|A)$
contours of predictive distribution
Dow Price/Dividend Ratio History

- Many applications posit a model, condition on data, and use the posterior.
- We can now apply this kind of data analysis to very large data sets.

Nonconjugate variational inference

- The class of conditionally conjugate models is very flexible.
- However, some models—like the CTM and DTM—do not fit in.
- In the past, researchers developed tailored optimization procedures for fitting the variational objective.
- We recently developed a more general approach that subsumes many of these strategies.

Nonconjugate variational inference

- Bishop (2006) showed that the optimal mean-field variational distribution is

$$
\begin{aligned}
q^*(z) &\propto \exp\left\{\mathrm{E}_{q(\beta)}\left[\log p(z \mid \beta, x)\right]\right\} \\
q^*(\beta) &\propto \exp\left\{\mathrm{E}_{q(z)}\left[\log p(\beta \mid z, x)\right]\right\}
\end{aligned}
$$

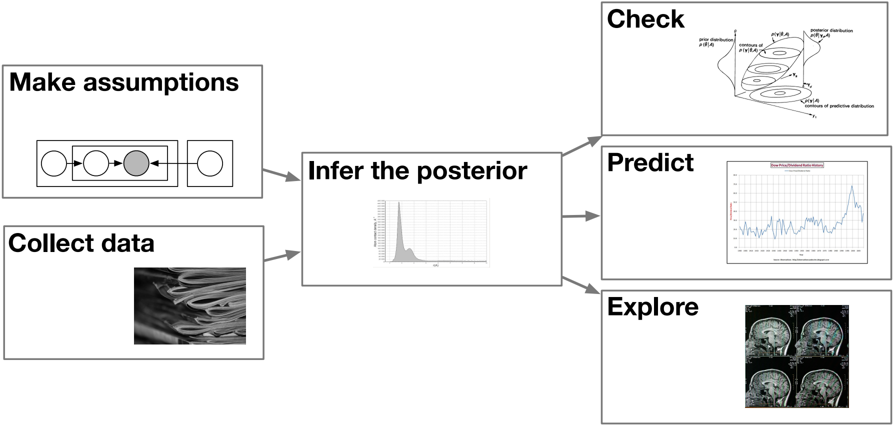
- In conjugate models, we can compute these expectations.
  This determines the form of the optimal variational distribution.

- In nonconjugate models we can't compute the expectations.

- But, under certain conditions, we can use Taylor approximations.
  This leads to Gaussian variational distributions.

Using and Checking Topic Models

Using and checking topic models



Make assumptions

Collect data

Infer the posterior

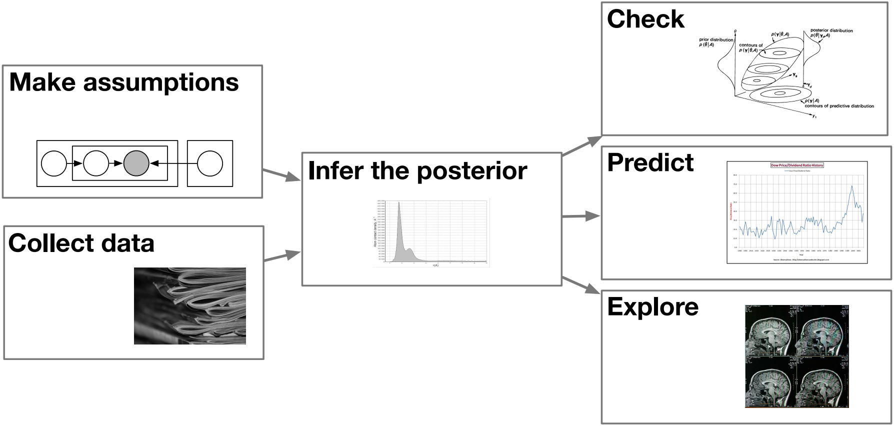
Check

Predict

Explore

- We have collected data, selected a model, and inferred the posterior.
- How do we use the topic model?

Using and checking topic models



Make assumptions

Collect data

Infer the posterior

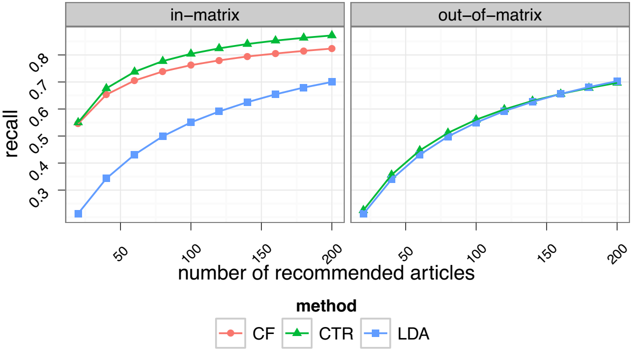
Check

Predict

Explore

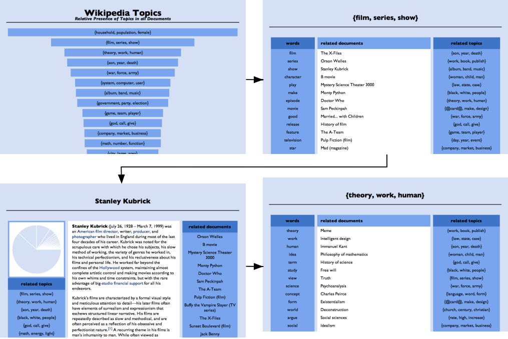
- Using a model means doing something with the posterior inference.
- E.g., visualization, prediction, assessing document similarity, using the representation in a downstream task (like IR)

Using and checking topic models



- Questions we ask when evaluating a model:
  - Does my model work? Is it better than another model?
  - Which topic model should I choose? Should I make a new one?
- These questions are tied up in the application at hand.
- Sometimes evaluation is straightforward, especially in prediction tasks.

Using and checking topic models



Wikipedia Topics

Relative Presence of Topics in all Documents

{household, population, female}

{film, series, show}

{theory, work, human}

{son, year, death}

{war, force, army}

{system, computer, user}

{album, band, music}

{government, party, election}

{game, team, player}

{god, call, give}

{company, market, business}

{math, number, function}

{city, large, area}

{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}

Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related topics
{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
{math, energy, light}

related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)
Jack Benny

{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

- But a promise of topic models is that they give good **exploratory tools.** Evaluation is complicated, e.g., is this a good navigator of my collection?
- And this leads to more questions:
  - How do I interpret a topic model?
  - What quantities help me understand what it says about the data?

Using and checking topic models

- How to interpret and evaluate topic models is an active area of research.
  - Visualizing topic models
  - Naming topics
  - Matching topic models to human judgements
  - Matching topic models to external ontologies
  - Computing held out likelihoods in different ways
- I will discuss two components:
  - **Predictive scores** for evaluating topic models
  - **Posterior predictive checks** for topic modeling

The predictive score

- Assess how well a model can predict **future data**
- In text, a natural setting is one where we observe part of a new document and want to predict the remainder.
- The **predictive distribution** is a distribution conditioned on the corpus and the partial document,

$$
\begin{aligned}
p(w \,|\, \mathscr{D}, \mathbf{w}_{\text{obs}}) &= \int_{\beta} \int_{\theta} \left(\textstyle\sum_{k=1}^{K} \theta_k \beta_{k,w} \right) p(\theta \,|\, \mathbf{w}_{\text{obs}}, \beta) p(\beta \,|\, \mathscr{D}) \\
&\approx \int_{\beta} \int_{\theta} \left(\textstyle\sum_{k=1}^{K} \theta_k \beta_{k,w} \right) q(\theta) q(\beta) \\
&= \mathrm{E}_q[\theta \,|\, \mathbf{w}_{\text{obs}}]^{\top} \mathrm{E}_q[\beta_{\cdot,w} \,|\, \mathscr{D}].
\end{aligned}
$$

The predictive score

- The **predictive score** evaluates the remainder of the document independently under this distribution.

$$s = \sum_{w \in \mathbf{w}_{\text{held out}}} \log p(w \,|\, \mathscr{D}, \mathbf{w}_{\text{obs}}) \tag{1}$$

- In the predictive distribution, *q* is any approximate poterior. This puts various models and inference procedures on the same scale.

- (In contrast, perplexity of entire held out documents requires different approximations for each inference method.)

The predictive score

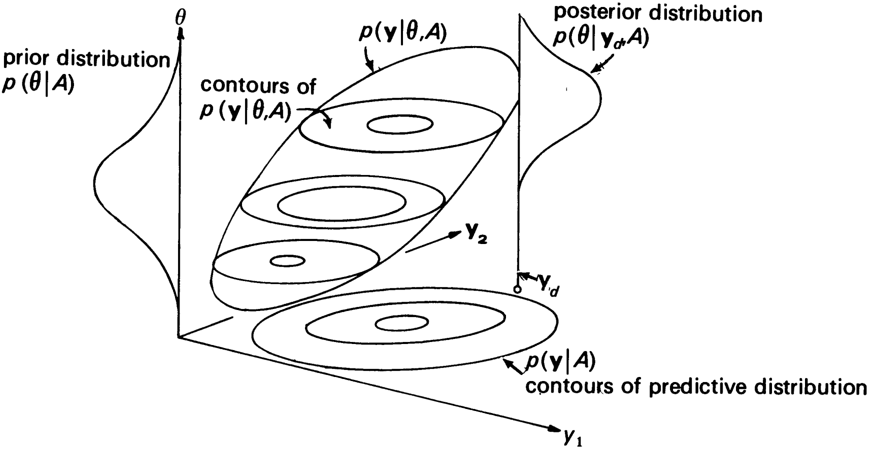
	Nature	*New York Times*	*Wikipedia*
LDA 100	-7.26	-7.66	-7.41
LDA 200	-7.50	-7.78	-7.64
LDA 300	-7.86	-7.98	-7.74
HDP	**-6.97**	**-7.38**	**-7.07**

The predictive score on large corpora using stochastic variational inference

Posterior predictive checks

- The predictive score and other model selection criteria are good for choosing among several models.
- But they don't help with the model building process; they don't tell us *how* a model is misfit. (E.g. should I go from LDA to a DTM or LDA to a CTM?)
- Further, prediction is not always important in exploratory or descriptive tasks. We may want models that capture other aspects of the data.
- **Posterior predictive checks** are a technique from Bayesian statistics that help with these issues.

Posterior predictive checks



This is a **predictive check** from Box (1980).

Posterior predictive checks

- Three stages to model building: estimation, criticism, and revision.
- In **criticism**, the model “confronts” our data.
- Suppose we observe a data set **y**. The predictive distribution is the distribution of data *if the model is true*:

$$p(y \mid M) = \int_\theta p(y \mid \theta) p(\theta)$$

- Locating **y** in the predictive distribution indicates if we can “trust” the model.
- Or, locating a **discrepancy function** $g(\mathbf{y})$ in its predictive distribution indicates if what is important to us is captured in the model.

Posterior predictive checks

- Rubin (1984) located the data **y** in the **posterior** $p(y \mid \mathbf{y}, M)$.
- Gelman, Meng, Stern (1996) expanded this idea to "realized discrepancies" that include **hidden variables** $g(\mathbf{y}, \mathbf{z})$.
- We might make modeling decisions based on a variety of simplifying considerations (e.g., algorithmic). But we can design the realized discrepancy function to capture what we really care about.
- Further, realized discrepancies let us consider which **parts of the model** fit well and which parts don't. This is apt in exploratory tasks.

Posterior predictive checks in topic models

- Consider a decomposition of a corpus into topics, i.e., $\{w_{d,n}, z_{d,n}\}$. Note that $z_{d,n}$ is a latent variable.

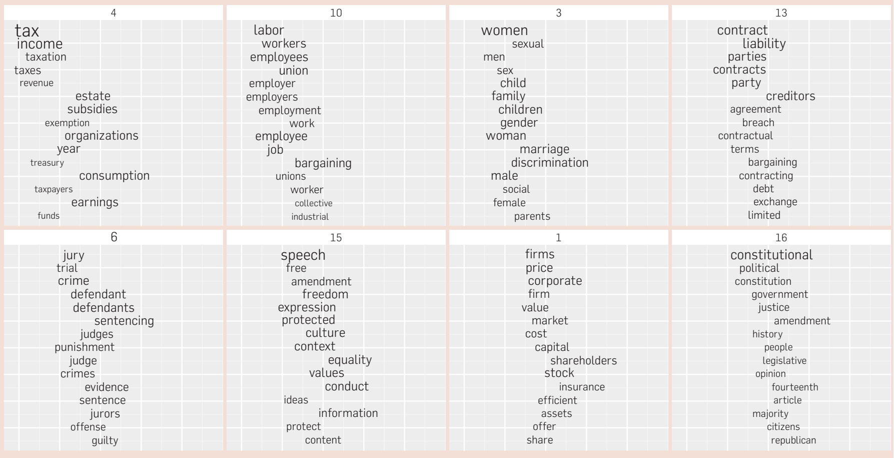
- For all the observations assigned to a topic, consider the variable $\{w_{d,n}, d\}$. This is the observed word and the document it appeared in.

- One measure of how well a topic model fits the LDA assumptions is to look at the **per-topic mutual information** between w and d.

- If the words from the topic are independently generated then we expect lower mutual information.

- What is "low"? To answer that, we can shuffle the words and recompute. This gives values of the MI when the words are independent.

Posterior predictive checks in topic models



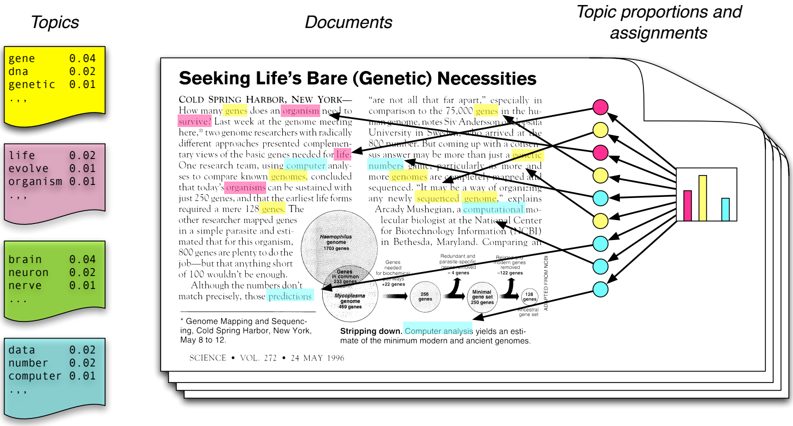
- This realized discrepancy measures model fitness
- Can use it to measure model fitness **per topic**.
- Helps us explore parts of the model that fit well.

Discussion

Probabilistic topic models

- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I evaluate and check a topic model?
- What are some unanswered questions in this field?
- How can I learn more?

Introduction to topic modeling



Topics

Documents

Topic proportions and assignments

gene 0.04
dna 0.02
genetic 0.01
. , ,

life 0.02
evolve 0.01
organism 0.01
. , ,

brain 0.04
neuron 0.02
nerve 0.01
. . .

data 0.02
number 0.02
computer 0.01
. , ,

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

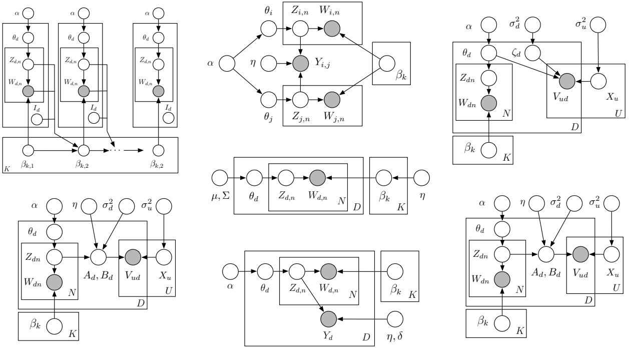
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

- LDA assumes that there are *K* topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

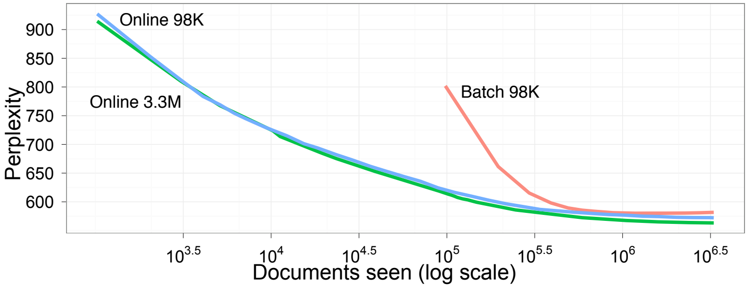
Extensions of LDA



Topic models can be adapted to many settings

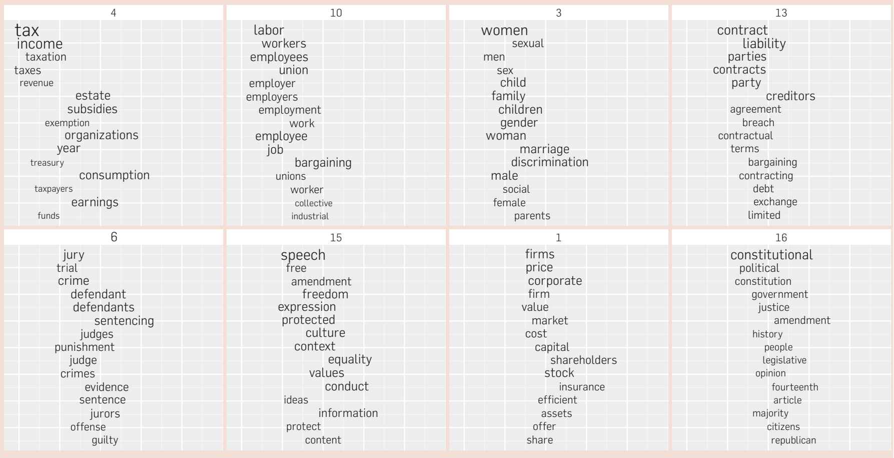
- relax assumptions
- combine models
- model more complex data

Posterior inference

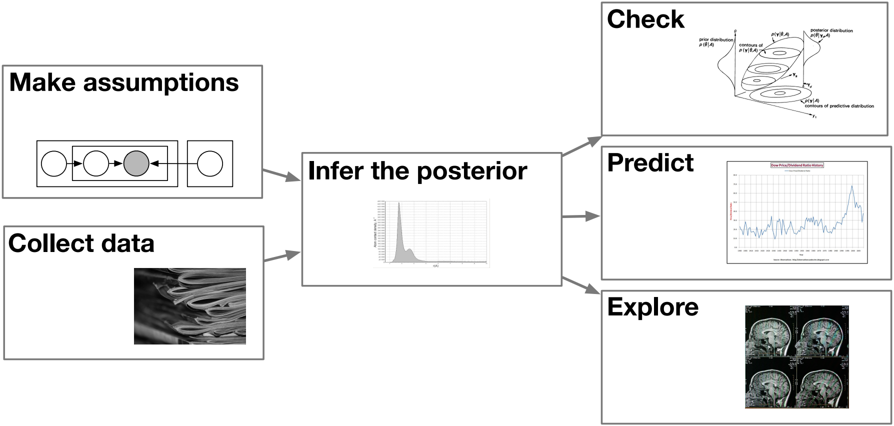


- Posterior inference is the central computational problem.
- Stochastic variational inference is a scalable algorithm.
- We can handle nonconjugacy with Laplace inference.
- (Note: There are many types of inference we didn't discuss.)

Posterior predictive checks



Probabilistic models



Make assumptions

Collect data

Infer the posterior

Check

Predict

Explore

Implementations of LDA

There are many available implementations of topic modeling.
Here is an incomplete list—

LDA-C*	A C implementation of LDA
HDP*	A C implementation of the HDP (“infinite LDA”)
Online LDA*	A python package for LDA on massive data
LDA in R*	Package in R for many topic models
LingPipe	Java toolkit for NLP and computational linguistics
Mallet	Java toolkit for statistical NLP
TMVE*	A python package to build browsers from topic models

* available at www.cs.princeton.edu/∼blei/

Research opportunities in topic modeling

- **New applications of topic modeling**
  What methods should we develop to solve problems in the computational social sciences? The digital humanties? Digital medical records?
- **Interfaces and downstream applications of topic modeling**
  What can I do with an annotated corpus? How can I incorporate latent variables into a user interface? How should I visualize a topic model?
- **Model interpretation and model checking**
  Which model should I choose for which task? What does the model tell me about my corpus?

Research opportunities in topic modeling

- **Incorporating corpus, discourse, or linguistic structure**
  How can our knowledge of language help inform better topic models?

- **Prediction from text**
  What is the best way to link topics to prediction?

- **Theoretical understanding of approximate inference**
  What do we know about variational inference? Can we analyze it from either the statistical or learning perspective? What are the relative advantages of the many inference methods?

- **And many specific problems**
  E.g., sensitivity to the vocabulary, modeling word contagion, modeling complex trends in dynamic models, robust topic modeling, combining graph models with relational models, ...

“We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.”

(J. Tukey, *The Future of Data Analysis*, 1962)

“Despite all the computations, you could just dance to the rock ’n’ roll station.”

(The Velvet Underground, *Rock & Roll*, 1969)